Efficient inference of cancer progression models

Supplementary Materials

Daniele Ramazzotti Giulio Caravagna Loes Olde Loohuis
Alex Graudenzi Ilya Korsunsky Giancarlo Mauri Marco Antoniotti
Bud Mishra

Monday 18th August, 2014

Keywords: *causality, cancer progression, prima facie, probability raising, temporal priority, direct acyclic graph, bootstrap, hypothesis testing, bayesian inference, leukemia, lung cancer.*

Acronyms: *Direct Acyclic Graph* (DAG), *Probability Raising* (PR), *Conjunctive Normal Form* (CNF), *Disjunctive Normal Form* (DNF), *Bayesian Network* (BN), *Cancer Progression Inference* (CAPRI).

Contents

List of Figures

Preamble

In what follows, we will use the notations, described below, in a manner consistent with the main body of the paper. *Atomic events*, in general, will be denoted by small Roman letters, such as a, b, c, ...; when it is clear from the context that the event in the model is, in fact, a genomic mutational event, we may refer to it directly using the standard biological nomenclature, e.g., BRCA1, BRCA2, etc. – it would be especially true, in the sections describing applications to real data. Formulas over events will be mostly denoted by Greek letters, and their logical connectives with the usual "and" ($\wedge$), "or" ($\vee$) and "negation" ($\bar{\cdot}$) symbols. Standard operations on sets will be used as well.

We will not employ distinct notations to denote *observed probabilities* and probabilities in the model which we aim at inferring (i.e. the *"theoretical probabilities"*). Which quantity is being referred to, will be made clear from the context. In the following, $\mathcal{P}(x)$ will denote the *probability* of x; $\mathcal{P}(x \wedge y)$, the *joint probability* of x *and* y, which will be naturally extended to the notation $\mathcal{P}(x \wedge y_1 \wedge \ldots \wedge y_n)$ for an arbitrary arity; and $\mathcal{P}(x \mid y)$, the *conditional probability* of x *given* y. Here x and y are formulas over events.

As with the discussion of causal structures in Section §1, we will write $c \rhd e$, where c and e are events being modeled, in order to denote the causal relation "c causes e". As we extend our presentation to general *formulas*, we will generalize the notation to $\varphi \rhd e$ with the meaning generalized *mutatis mutandis*[1].

Our Supplementary Materials are structured as follows: §1 and §2 introduce and comparatively study, without any pretense of exhaustivity, the current state-of-the-art in "'causation theories", and Bayesian networks inference; §3 next introduces the algorithmic framework and foundation for §4; finally, §5 and §6 conclude with analyses of experimental results. The goal of the Supplemental Materials is to present to a wide multi-disciplinary audience sufficient amount of details about both the theories and causality algorithms (with proper citations to the implementations) in order that s/he is able to reproduce and verify our results, as described in the main body of the paper.

1 Foundations of causation

In this section, we start with an outline of the current state-of-the-art theories of causation, which enjoys a long and colorful history, starting with the work of Avicenna circa 1000 AD. However, we restrict our description only to the main ideas and limitations of these theories, as a more detailed discussion of various topics related to these theories is available elsewhere (see [1] or [2]).

Our biological notion of causality is firmly grounded on the notions of *Darwinian evolution*: in that, it is about an *ensemble of entities* (e.g., population of cells, organisms, etc.). Within this ensemble, a causal event (say c) in a member entity may result in variations (changes in genotypic frequencies); such variations are exhibited in the phenotypic variations within the population, which is subject to Darwinian positive (and subsequently, Malthusian negative) selections, and sets the stage for a new effect event (say e) to be selected, should it occur next; we then conclude that "$c \rhd e$." For an example of how to interpret EGFR $\rhd$ CDK, see the introduction of the main text.

While there could be other meaningful extensions of this framework (see [3])[2] , we believe that

[1] Note that the scope of this study is intentionally kept limited from further generalizing the "causal formulas"; for instance, we will not deal with any example of the form $\varphi_i \rhd \varphi_j$, where $\varphi.$ could be any general formula (including a complex causal formula or a temporal formula). This choice is justified in view of complexity, practicality, applicability and expressiveness in the context of cancer progression driven by somatic evolution.

[2] Also see, the debate between Fisher and Wright, in response to Fisher's *fundamental theorem of genetics*.

it suffices in describing the causality relations implicit in the somatic evolution responsible for tumor progression. Note further that by its very statistical nature, we capture just those relations that only reflect *"Type-level Causality"*, and relegate *"Token-level Causality"*, – a more nuanced concept – to the future research. Thus, note that, while we can estimate, for a population of cancer patients of a particular kind (say atypical Chronic Myeloid Leukemia, aCML, patients) whether and with what probability a mutation (such as SETBP1) would cause certain other mutations (such as ASXL1 single nucleotide variants or *in-del*) to occur, it will remain silent as to whether a particular ASXL1 mutation in a particular patient was caused by an earlier SETBP1 mutation.

Based on the afore-mentioned biological framework, we will focus primarily on how to devise efficient and accurate algorithms for extracting causal relations from the patient genomic data; we leave it to the readers to intuit how an inferred causal relation may be verified/refuted by *in vitro* or *in silico* experiments and how it could be used in therapy design that would guide the clocks involved in cancer's natural somatic evolution (more details are forthcoming).

1.1 Hume's regularity theory

The modern study of causation begins with the Scottish philosopher David Hume (1711-1776). According to Hume, a theory of causation could be defined axiomatically, using the following ingredients: *temporal priority*, implying that causes are invariably followed by their effects [4], augmented by various constraints, such as contiguity, constant conjunction[3], etc. Theories of this kind, that try to analyze causation in terms of invariable patterns of succession, have been referred to as *regularity theories* of causation.

Nonetheless, the notion of causation has spawned far too many variants and has been a source of acerbic debates. All these theories present well-known limitations and confusion, but have led to a small number of modern versions of commonly accepted (at least among the philosophers) frameworks. See the theories discussed and studied by Suppes et al. §1.2, Lewis et al. §1.3, and Pearl et al. §1.3. One of the most prominent among these is Suppes' *probabilistic* causation, whose axioms are expressible in *probabilistic propositional modal logics*, and amenable to algorithmic analysis. It is the framework upon which we build our analyses and algorithms.

We will momentarily discuss the main limitations of regularity theories [2], in order to better prepare the reader for the subsequent discussions of these theories and the algorithms to which they lead. Thus, the next three sections will focus on two issues: (*i*) how the state-of-the-art theories of causation have attempted formulating a *sound and complete* theory of causation, as well as (*ii*) what unsolved problems in this framework still remain open.

Imperfect regularities. In general, we cannot state that causes are *invariably* (i.e., without fail) followed by their effects. For example, while we may state that "smoking is a cause of lung cancer", we do grant that there would be still some smokers who do not develop lung cancer.

Situations such as these are referred to as *imperfect regularities*, and could arise for many different reasons. One of these – which is a very common situation in the context of cancer – involves the heterogeneity of the situations in which a cause resides. For example, some smokers may have a genetic susceptibility to lung cancer, while others do not; moreover, some non-smokers may be exposed to other carcinogens, while others are not. Thus, the fact that not all smokers develop lung cancer can be explained in these terms.

[3]Some of these notions have been modernized with the introduction of the machinery from statistical inference, logic and model theory; but they have stayed more or less true to Hume's programme.

Irrelevance. An event that is invariably followed by another, can be irrelevant to it. Consider the example in [5]: salt that has been hexed by a sorceror invariably dissolves when placed in water, but hexing does not cause the salt to dissolve. In fact, hexing is irrelevant for this outcome. Probabilistic theories of causation capture exactly this situation by requiring that causes alter the probabilities of their effects, see §1.2.

Asymmetry. If we claim that an event c causes another event e, then, typically, we would anticipate being able to claim that e does not cause c, which would naturally follow from a strict temporal-priority-constraint: *cause precedes effect temporally.* In the context of the preceding example, smoking causes lung cancer, but lung cancer does not cause one to smoke.

Spurious regularities. Consider a situation – not very uncommon – where a unique cause is regularly followed by two or more effects. As an example, suppose that one observes the height of the column of mercury in a particular barometer dropping below a certain level. Shortly afterwards, because of the drop in atmospheric pressure (the unobserved cause for falling barometer), a storm occurs. In this settings, a regularity theory could claim that the drop of the mercury column causes the storm when, indeed, it is only correlated to it. Following common terminologies, we will say that such situations are due to *spurious correlations.* There now exists an extensive literature discussing such subtleties that are important in understanding the philosophical foundations of causality theory; see [2].

1.2 Probabilistic theories of causation

In this section we will introduce the notion of *probabilistic causation.* The basic idea behind these theories is that "causes alter the probabilities of their effects;" see [2] for details.

Suppes' prima facie cause

Patrick Suppes proposed the notion of a *prima facie cause* that represents the core of *probabilistic causation* and also provides the algorithmic foundations of our analysis.

Definition 1 (Probabilistic causation, [6])**.** *For any two events c and e, occurring respectively at times t_c and t_e, under the mild assumptions that $0 < \mathcal{P}(c), \mathcal{P}(e) < 1$, the event c is called a* prima facie cause *of e if it occurs* before *and* raises the probability *of e, i.e.*

$$t_c < t_e \quad \textit{and} \quad \mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c})\,. \tag{1}$$

From now on, the former condition will be referred to as *temporal priority*, whereas the latter as *probability raising* (PR). This notion of causation has some advantages over the simplest version of a regularity theory of causation, e.g., it deals with various issues usually associated with imperfect regularities (§1.1).

Unfortunately, however, prima facie causality is still *not sufficient* in capturing a causation relationship *in its full generality.* For instance, the problem of spurious regularities still remains, additionally requiring that prima facie causes be refined further into two classes: *genuine* and *spurious.* In the latter case, as discussed, we may observe a prima facie cause to be so labeled only because of spurious correlations. Also, as discussed extensively in the literature, one may encounter certain situations, in which Suppes' characterization fails to provide a *necessary* condition. In the next two paragraphs, we will briefly discuss an attempt to make Suppes' conditions sufficient for any causal claims, and another to determine when it is not necessary.

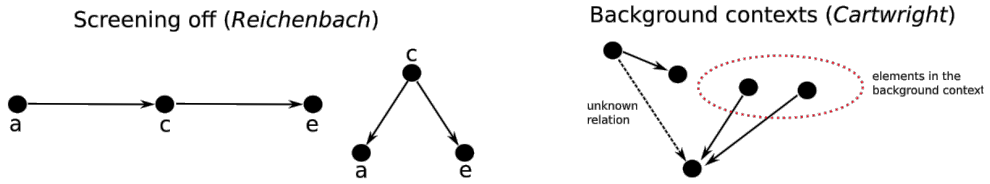


Figure 1: **Example of screening-off and of background context.** (*left*) Example of Reichenbach's screening-off where c is a genuine cause of e and a is a genuine cause of c, and the correlations between a and e are only just manifestations of these known causal connections, and c is a common cause of both a and e, that is exactly the situation of spurious correlation described in §1.1. (*right*) Example of Cartwright's background context.

Reichenbach's screening-off

In [7], Reichenbach discussed the notion of *screening-off* to describe a particular type of probabilistic relationship. Consider, e.g., events a, c and e, and assume to observe $\mathcal{P}(e \mid a \wedge c) = \mathcal{P}(e \mid c)$, then we say that c is *screening* a *off* from e. When $\mathcal{P}(e \wedge c) > 0$, this is equivalent to stating that $\mathcal{P}(a \wedge e \mid c) = \mathcal{P}(a \mid c) \cdot \mathcal{P}(e \mid c)$ – i.e., a and e happen to be probabilistically independent, when conditioned upon c. The preceding situation could occur in two cases, see Figure §1.

In the first case, c is a genuine cause of e while a is a genuine cause of c as well, and the correlations between a and e are only just manifestations of these known causal connections. For example, unprotected sex (a) appears to cause AIDS (e) *only* because of sexually transmitted HIV infection (c). Then, we would expect that among those who have already been infected with HIV, the probability of contacting AIDS would be the same regardless of whether one is engaged in unprotected sex or not. Here c is a *proximate* cause of e and an *intermediate* cause leading from a to e, i.e. an instance of *causal transitivity*. In the second case, c is a common cause of both a and e, that is exactly the situation of spurious correlation described in §1.1.

Building upon this idea, Reichenbach formulated the *Common Cause Principle* (CCP) to detect situations leading to "screening-off," and so identify when a spurious correlation can be explained in terms of a common cause. Unfortunately, there are situations where such a principle leads to computationally intractable criteria. Since, these issues are not germane to our context, we will not discuss them further, other than pointing the interested readers to appropriate literature [2]. Nevertheless, the idea of screening-off has significantly influenced some of the most widely-used recent theories of causation, and has become central to the topic.

Simpson's paradox and Cartwright's background context

Up to now, we have discussed the *sufficiency* (or lack of it) of the characterization for causality provided in the Reichenbach-Suppes framework. Conversely, we may also examine those situations where this framework also fails to give all the *necessary* conditions for a causal claim. For example, consider smoking as a cause of lung cancer. But, examine in details a situation where it so happens that smoking is highly correlated with living in the country: those who live in the country are much more likely to smoke than those who do not. Suppose now that city pollution is a second cause of lung cancer, which happens to be a much stronger cause than smoking. Consider now the problem of causal claims on the combination of these two heterogenous populations: including those who live in the country and those who do not. Then, an analysis of

those two populations in combination may falsely lead to the conclusion that smokers are, over all, less likely to suffer from lung cancer than non-smokers. This example is an instance of the so-called *Simpson's Paradox*, which has been discussed extensively by various philosophers (see Nancy Cartwright [8] and Brian Skyrms [9]).

Cartwright and Skyrms introduced the concept of *background contexts* to explain and correct this problem. Let us call the set of all the factors that are causes of the event e (a factor can be an *atomic* event but it can also be *the composition of a set* of events), but are not caused by the event c, the set of *independent causes* of e. A background context for a causal relationship from c to e is the maximal conjunction of factors, each of which is either an independent cause of e, or the negation of an independent cause of e (as shown in Figure §1). We will denote by variables $b_1, \ldots, b_n$ all the background contexts of a causal relationship. According to Cartwright then, c causes e if and only if $\mathcal{P}(e \mid c \wedge b_i) > \mathcal{P}(e \mid \overline{c} \wedge b_i)$, that is if c raises the probability of e in every background context $b_i \in B$. Skyrms proposed a slightly weaker condition: a cause must raise the probability of its effect in at least one background context, without lowering it in any other.

Eells' taxonomy

Cartwright defined a cause in terms of raising the probability of its effect. But there are other possible probabilistic relations between c and e, as described, for instance, by Eells, who proposes the following taxonomy [10]: (i) c is a *cause* of e if and only if it raises its probability in every background context B, (ii) c is an *inhibition* for e when it lowers such a probability, (iii) c is *causally irrelevant* to e when it does not change it and, finally, (iv) c is a *mixed cause* of e, otherwise.

This supplemental material (SM) will adhere to the basic idea of a cause being a *probability-raiser* of its effect and ignore for the time being all other variants. According to Suppes' probabilistic theories of causation, we can evaluate a causal claim in terms of Definition §1, further augmented by the ideas of screening-off and background contexts; the same algorithmic, inferential and logical tools that we propose here can be used *mutatis mutandis*, should a user wish to explore a variant framework leading to a different axiomatic formulation of causation – provided its expressivity is limited to a probabilistic propositional modal logic – as seems the case to be.

Issues of probabilistic causation

Next we describe some thorny issues in the theory of probabilistic causation. We also briefly point out some unresolved problems, proposed plans of attack, and ensuing criticisms. For a deeper discussion see [2].

Pearl's criticism. In [11], Pearl argues that the notion that causes "raise the probabilities" of their effects *cannot be expressed in the language of probability theory*. In particular, according to Pearl, the inequality $\mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c})$ fails to capture the intuition behind probability raising, which must be *manipulative* or *counterfactual*. Because of this limit, Pearl argues that it is not possible to rigorously describe the intuitions behind the probability raising theory and, for this reason, the only way to properly assess a causal claim is exclusively by *intervention*. The methods described in this supplemental material (SM) are not negated by these arguments as our model reasons about an ensemble (tumor with heterogeneous cell-types) and type-level causality, expressed in a powerful language of probabilistic modal logic. Pearl's theory is discussed further in §1.3.

Determining the background context. As described, the background contexts of a claim are all the factors causally relevant to the effect, but not to the cause. This assumption appears

to prevent Cartwright's theory from being a reductive analysis of causation. In fact, the theory appeals to causal relations to define a set of probabilistic constraints on the possible causal claims compatible with the observations in terms of probabilities. In any case, even if there is no reduction of causation to probability, in practice, it can be difficult (or algorithmically complex) to determine the background contexts without knowing the causal topology in advance. Unfortunately, this argument introduces an unavoidable *circularity*.

1.3 Counterfactual theories of causation

Next we briefly discuss *counterfactual* theories of causation where the meaning of causal claims is explained in terms of a *possible-world semantics* and counterfactual conditionals of the form: *had c not occurred, e would not have occurred either*. For detailed discussions see [12].

Lewis's counterfactuals

The most complete known counterfactual theory of causation is due to David Lewis [13] and exploits a possible world semantics to state truth conditions for counterfactuals in terms of *similarity* among possible worlds: one possible world is closer to actuality than another, if it is more similar to the actual world.

Following this idea, Lewis defined two important constraints on the resulting similarity relation: (i) similarity induces an ordering of worlds in terms of closeness to the actual world and (ii) the actual world is the closest possible world to actuality. Then, the evaluation of the counterfactual *"if c were the case, e would be the case"* is true just in case it is closer to actuality to make the first term true along with the second – as opposed to making it true without. Therefore, in terms of counterfactuals Lewis defines the following notion of causality: given c and e, whether e occurs or not depends on whether c occurs or not, and e causally depends on c if and only if, if c were not to occur e would not occur. Thus, the idea of cause is conceptually linked to the idea of *something that makes a difference,* and this concept in turn is naturally described in terms of counterfactuals. Lewis also characterized causation in terms of temporal direction by stating that the direction of causation is the direction of causal dependence and that, typically, events causally depend on earlier events but not on later ones.

Causal Chains. In [13], Lewis states that causal dependence between events is *sufficient but not necessary*, i.e., it is possible to have causation without causal dependence. Consider, e.g., when c causes d, which in turn causes e; Lewis argues that c must cause e as well by means of a transitivity. However, since causal dependence is not transitive as would be the case for causation according to Suppes, the causal relation between c and e may not be evident. To overcome this problem, Lewis defines a causal chain as the finite sequence of events c, d and e and defines that *c is a cause of e if and only if there exists a causal chain leading from c to e.*

Issues of counterfactual causation

We briefly describe some issues inherent to these theories; for a deeper discussion, see [12].

Context-sensitivity. Lewis's theory assumes that causation is an absolute relation, whose nature does not vary from one context to another. This approach has recently been criticized since it often leads to absurd results [12], as demonstrated by various easy-to-construct counter-examples.

Transitivity and Preemption. As discussed above, Lewis incorporates transitivities in his notion of causation by defining them in terms of chains of causal dependence. The transitivity of causation is sound in some contexts, but a number of counter-examples has been shown to cast doubts on this interpretation of causation [12]; the debate surrounding the transitivity of causation is unlikely to be easily settled. Nevertheless, in this work we aim at inferring *minimal models of causation*, in which each cause is sufficient for its child to occur. For this reason, we have opted to remove transitivities.

Manipulability theories of causation

We now briefly discuss the notion of *intervention* as propounded by Judea Pearl [11]; in general interventionist versions of manipulability theories can be seen as counterfactual theories. For a detailed discussion on this and manipulability theories of causation refer to [14].

Pearl characterizes his notion of intervention in terms of a primitive notion of causal mechanism. According to him, the world is organized in the form of stable mechanisms (i.e. physical laws) which are autonomous. Therefore, he states that we can change one of them, without changing all the others. Thus an intervention may imply that: *if we manipulate c and nothing happens, then c cannot be cause of e, but if a manipulation of c leads to a change in e, then we know that c is a cause of e, although there might be other causes as well.*

In other words, when among many events a causal relationship between some e and its parents (i.e. directed causes, say $c_1, \ldots, c_n$) is present, the interventions will disrupt completely the relationships between e and $c_1, \ldots, c_n$ such that the value of e is determined by the intervention only. Thus, intervention is a surgical operation in the sense that no other causal relationship in the system are changed by it. Hence, Pearl's assumption is that the other variables that change in values under this intervention will do so *only because they are effects of e*. Going back to the barometer example of §1.1: observing the drop of the mercury column increases the probability of a storm coming, but if we manipulate the drop of the mercury column by intervention such that its drop is caused by the intervention only, then we will be able to qualify barometer as a cause of storms instead of the drop itself. Pearl's theory has been very influential among the computational causality theorists, and has generated state-of-the-art algorithms for causal network inference, which we shortly present in §2 and use it as a benchmark to compare against; see §5.

Issues of interventionist causation

Next, we point the reader to some problems that can arise in practice, when applying intervention in the context of causal inference. For a deeper discussion we refer to [14].

Circularity. An intervention on an event e leaves intact *all* the other *causal* mechanisms besides the ones involving c as a cause. Because of this, Pearl's intervention could lead to circularity problems, i.e., it seems that the causal mechanisms need to be known in advance in order to asses them.

Possible and impossible interventions. Causal claims are described in terms of counterfactuals of what would happen when applying intervention to a given causal relationship. Moreover, the notion of intervention is connected with the possibility of a *human action* to intervene in a system. In some contexts, however, it may be *impossible* to evaluate what would happen by performing a *surgical* intervention. Thus, it should be clear that, regardless of the possible criticisms to Pearl's framework, there are situations where, at least relative to the current human capabilities, it is very complicated, if not impossible, to perform intervention.

1.4 Our simplified framework

It should be clear that the currently existing literature lacks a framework readily applicable to the problem of reconstructing cancer progression, as governed by somatic evolution; however, each theory has ingredients that are highly promising and relevant to the problem.

Each of the existing theories faces various difficulties, which are rooted primarily in the attempt to construct a framework in its full generality: *each theory aims to be both necessary and sufficient for any causal claim, in any context.* In contrast, this supplemental material (SM) simplifies the problem by breaking the task into two: first, define a framework for Suppes' prima facie notion though it admits some spurious causes (§3), but then deal with spuriousness by using a combination of tools, e.g., Bayesian, empirical Bayesian, regularization, which we recall in §2. The framework is based on a set of conditions that are *necessary even though not sufficient* for a causal claim, and is used to refine a prima facie cause to either a genuine or a spurious cause (or even ambiguous ones, to be treated as plausible hypotheses which can be refuted/validated by other means).

Statement of assumptions. Along with the described interpretation of causality, through out this document, we make following simplifying assumptions:

(*i*) *All causes involved in cancer can be expressed by monotonic Boolean formulas*: i.e., all causes are positive and can be expressed in CNF where all literals occur only positively. The size of the formula and each clause therein are bounded by small constants.

(*ii*) *All events are persistent*: i.e., once a mutation has occurred, it cannot disappear. Hence, we do not model situations where $\mathcal{P}(e \mid c) < \mathcal{P}(e \mid \overline{c})$.

(*iii*) *Closed world*: all the events which are causally relevant for the progression are observable and the observation can significantly describe the progressive phenomenon.

(*iv*) *Relevance to the progression*: all the events have probability strictly in the real open interval $(0, 1)$, i.e. it is possible to asses if they are relevant to the progression.

(*v*) *Distinguishability*: no two events appear equivalent, i.e. they are neither both observed nor both missing *simultaneously.*

2 Structural learning of Bayesian Networks (BNs)

In this section we briefly discuss the notion of *Bayesian Network* (BN) and how to learn both its parameters and structure *ab initio*, with no prior knowledge. For a detailed discussion on the topic, refer to [15, 16]. This section is intended to be accessible to a non-technical audience, although citations are provided for technical resources on each algorithm discussed.

2.1 Preliminaries

A BN is a statistical model that succinctly represents a *joint distribution* over n variables and encodes it in a *direct acyclic graph* over n nodes (one per variable)[4]. In BNs, the full joint distribution can be written as a product of conditional distributions on each variable. An edge between two nodes A and B denotes statistical dependence, $\mathcal{P}(A \wedge B) \neq \mathcal{P}(A)\mathcal{P}(B)$, no matter on which other variables we condition on (i.e., for any other set of variables $\mathcal{C}$ it holds

[4]In our setting, each variable is a modeled event and, for consistency with the BN notation, we will denote these as capital letters in this section.

$\mathcal{P}(A \wedge B \mid \mathcal{C}) \neq \mathcal{P}(A \mid \mathcal{C})\mathcal{P}(B \mid \mathcal{C})$. In such a graph, the set of variables connected to a node X determines its set of "parent" nodes $\pi(X)$. Note that a node cannot be both ancestor and descendant of another node, as this would cause a directed cycle.

Finally, the joint distribution over all the variables can be written as $\prod_X \mathcal{P}(X \mid \pi(X))$. Of course, if a node has no incoming edges (i.e. no parents), we simply use its marginal probability $\mathcal{P}(X)$. Thus, to compute the probability of any combination of values over the variables, we need only parameterize the conditional probabilities of each variable given its parents. If the variables are binary, the number of parameters in each conditional probability table is locally of exponential size: namely, $2^{|\pi(X)|} - 1$. Thus, the total number of parameters needed to compute the full joint distribution is only of size $\sum_X 2^{|\pi(X)|} - 1$, which is considerably less than $2^n - 1$.

A useful property of the graph structure is that we can define, for each variable, a set of nodes called the *Markov blanket* so that, conditioned on it, this variable is independent of all other variables in the system. It can be proven that for any BN, the Markov blanket consists of a node's parents, children as well as the parents of the children.

The usage of the symmetrical notion of conditional dependence introduces important limitations of structure learning in BNs. In fact, note that edges $A \rightarrow B$ and $B \rightarrow A$ denote equivalent dependence between A and B, thus distinct graphs model the exact same set of independence and conditional independence relations. This yields the notion of *Markov equivalence class* as a *partially directed acyclic graph*, in which the edges that can take either orientation are left undirected. A theorem proves that two BNs are Markov equivalent when they have the same *skeleton* and the same *v-structures*, the former being the set of edges, ignoring their direction (e.g., $A \rightarrow B$ and $B \rightarrow A$ constitute a unique edge in the skeleton) and the latter being all the edge structures in which a variable has at least two parents, but those do not share an edge (e.g., $A \rightarrow B \leftarrow C$)[5] [17].

BNs have an interesting relation to canonical boolean logical operators $\wedge$, $\vee$ and $\oplus$ and formulas over variables. In fact these formulas, which are "deterministic" in principle, in BNs are naturally softened into *probabilistic relations* to allow some degree of uncertainty or noise. This probabilistic approach to modeling logic allows representation of qualitative relationships among variables in a way that is inherently robust to small perturbations by noise. For instance, the phrase *"in order to hear music when listening to an mp3, it is necessary and sufficient that the power is on and the headphones are plugged in"* can be represented by a probabilistic conjunctive formulation that relates power, headphones and music, in which the probability that music is audible depends only on whether power and headphones are present. On the other hand, there is a small probability that the music will still not play (perhaps we forgot to load any songs into the device) even if both power and headphones are on, and there is small probability that we will hear music even without power or headphone (perhaps we are next to a concert and overhear that music).

Note that in this review, we only consider the subset of networks that have discrete random variables that are visible. Networks with latent and continuous variables present their own challenges, although they share most of the mathematical foundations discussed here.

2.2 Approaches to learn the structure of a BN

Classically, there have been two families of methods aimed at learning the structure of a BN from data. The methods belonging to the first family seek to explicitly *capture all the conditional independence relations* encoded in the edges, and will be referred to as *constraint based approaches* (§2.2.1). The second family, that of *score based approaches* (§2.2.2), seeks to choose a model

[5] In BN terminology, parent A and C are considered "unwed parents." For this reason, the v-structure is often called an immorality or an *unshielded collider*.

that *maximizes the likelihood of the data* given the model. Since both the approaches lead to intractability (NP-hardness) [18, 19], computing and verifying an optimal solution is impractical and, therefore, heuristic algorithms have to be used, which only sometimes guarantee optimality. Recently, a third class of learning algorithms that takes advantage of *specialized logical relations* (mentioned in the previous section) have been introduced (§2.2.3). In the rest of this section we describe in detail some of these approaches. After our approach is introduced, we will compare its performance with that of all the techniques described below in §5.

2.2.1 Constraint based approaches

We present an intuitive explanation of several common algorithms used for structure discovery by explicitly considering conditional independence relations between variables. For more detailed explanations and analyses of complexity, correctness and stability, refer to the related references.

The basic idea behind all algorithms is to build a graph structure reflecting the independence relations in the observed data, thus matching as closely as possible the empirical distribution. The difficulty in this approach lies in the number of conditional pairwise independence tests that an algorithm would have to perform to test all possible relations. This is indeed *exponential* requiring to condition on a power set, when testing for the conditional independence between two variables. This inherent intractability requires the introduction of *approximations*.

Here, we focus on two specific constraint based algorithms, the *PC algorithm* [20] and the *Incremental Association Markov Blanket* (IAMB, [21]), because of their proven efficiency and widespread usage. In particular, the PC algorithm solves the aforementioned approximation problem by conditioning on incrementally larger sets of variables, such that most sets of variables will never have to be tested, whereas the IAMB first computes the Markov blanket of all the variables and conditions only on members of the blankets. A few more details about these algorithms follow.

The PC algorithm. The PC algorithm [20] begins with a fully connected graph and, on the basis of pairwise independence tests, iteratively removes all the extraneous edges. It is based on the idea that if a separating set exists that makes two variables independent, we can remove the edge between them. To avoid an exhaustive search of separating sets, these are ordered to find the correct ones early in the search. Once a separating set is found, the search for that pair can end. The PC algorithm orders separating sets of increasing size l starting from 0, the empty set, and incrementing until $l = n - 2$. The algorithm stops when every variable has fewer than $l - 1$ neighbors, since it can be proven that all valid sets must have already been chosen. During the computation, the larger the value of l is, the larger number of separating sets must be considered. However, by the time l gets too large, the number of nodes with degree l or higher must have dwindled considerably. Thus, in practice, we need only consider a small subset of all the possible separating sets.

Incremental Association Markov Blanket algorithm. A distinct type of constraint based learning algorithms uses the Markov blankets to restrict the subset of variables to test for independence. Thus, when this knowledge is available in advance, we do not have to test a conditioning on all possible variables. A widely used and efficient algorithm for Markov blanket discovery is IAMB. In it, for each variable X, we keep track of a hypothesis set $\mathcal{H}(X)$. The goal is for $\mathcal{H}(X)$ to equal the Markov blanket of X, $\mathcal{B}(X)$, at the end of the algorithm. IAMB consists of a forward and a backward phase. During the forward phase, it adds all possible variables into $\mathcal{H}(X)$ that could be in $\mathcal{B}(X)$. In the backward phase, it eliminates all the false positive variables from the hypotheses set, leaving the true $\mathcal{B}(X)$. The forward phase begins with an empty $\mathcal{H}(X)$ for

each X. Iteratively, variables with a strong association with X (conditioned on all the variables in $\mathcal{H}(X)$) are added to the hypotheses set. This association can be measured by a variety of non-negative functions, such as *mutual information*. As $\mathcal{H}(X)$ grows large enough to include $\mathcal{B}(X)$, the other variables in the network will have very little association with X, conditioned on $\mathcal{H}(X)$. At this point, the forward phase is complete. The backward phase starts with $\mathcal{H}(X)$ that contains $\mathcal{B}(X)$ and false positives, which will have little conditional association, while true positives will associate strongly. Using this test, the backward phase is able to remove the false positives iteratively until all but the true positives are eliminated.

2.2.2 Score based approaches

This approach to structural learning seeks to maximize the likelihood of a set of observed data. Since we assume that the data are independent and identically distributed, the likelihood of the data $\mathcal{L}(\cdot)$ is simply the product of the probability of each observation. That is,

$$\mathcal{L}(D) = \prod_{d \in D} \mathcal{P}(d)$$

for a set of observations D. Since we want to infer a model $\mathcal{G}$ that best explains the observed data, we define the likelihood of observing the data given a specific model $\mathcal{G}$ as

$$\mathcal{LL}(\mathcal{G}, D) = \prod_{d \in D} \mathcal{P}(d \mid \mathcal{G}) \,.$$

The actual likelihood is not used in practice, as this quantity becomes very small and impossible to represent in a computer. Instead, the logarithm of the likelihood is used for three reasons. First, the $log(\cdot)$ function is monotonic. Second, the values that the log-likelihood takes do not cause the same numerical problems that likelihood does. Third, it is easy to compute because the log of a product is simply the sum of the logs (e.g., $\log(xy) = \log x + \log y$), and the likelihood for a Bayesian network is a product of simple terms.

Practically, however, there is a problem in learning the network structure by maximizing log-likelihood alone. Namely, for any arbitrary set of data, the most likely graph is always the fully connected one (i.e. all edges are present), since adding an edge can only increase the likelihood of the data. To correct for this phenomenon, log-likelihood is almost always supplemented with a *regularization term* that penalizes the complexity of the model[6]. There are a plethora of regularization terms, some based on information theory and others on Bayesian statistics (see [22] and references therein), which all serve to promote *sparsity* in the learned graph structure, though different regularization terms are better suited for particular applications.

Also in this case we choose to describe a particularly relevant and known score, the *Bayesian Information Criterion* (BIC, [15]), which will be subsequently compared to the performance of our approach.

The Bayesian Information Criterion. BIC uses a score that consists of a log-likelihood term and a regularization term depending on a model $\mathcal{G}$ and data D

$$\text{BIC}(\mathcal{G}, D) = \mathcal{LL}(\mathcal{G}, D) - \frac{\log m}{2} \dim(\mathcal{G}). \tag{2}$$

[6]Note that more edges in a graph require more parameters in the conditional probability distributions, thus increasing model complexity. If it was known that the number of parameters for each node is fixed, then regularization is not necessary.

Here, D denotes the data, m denotes the number of samples and $\dim(\mathcal{G})$ denotes the number of parameters in the model. Because $\dim(\cdot)$ depends on the number of parents each node has, it is a good metric for model complexity. Moreover, each edge added to $\mathcal{G}$ increases model complexity. Thus, the regularization term based on $\dim(\cdot)$ favors graphs with fewer edges and, more specifically, fewer parents for each node. The term $\log m/2$ essentially weighs the regularization term. The effect is that the higher the weight, the more sparsity will be favored over "explaining" the data through maximum likelihood.

Note that the likelihood is implicitly weighted by the number of data points, since each point contributes to the score. As the sample size increases, both the weight of the regularization term and the "weight" of the likelihood increase. However, the weight of the likelihood increases faster than that of the regularization term[7]. Thus, with more data, likelihood will contribute more to the score, and we may trust our observations more and have less need for regularization. Statistically speaking, BIC is a *consistent score* [15]. In terms of structure learning, this observation implies that for sufficiently large sample sizes, the network with the maximum BIC score is *I-equivalent* to the true structure. Consequently, $\mathcal{G}$ contains the same independence relations as those implied by the true structure. As the independence relations are encoded in the edges of the graph, we are guaranteed to learn a Markov-equivalent network, with the same skeleton and the same v-structures as the true graph, though not necessarily with the correct orientations for each edge.

2.2.3 Learning logically constrained networks

In §2.1, we noted that an important class of BNs captures common binary logical operators, such as $\wedge$, $\vee$, and $\oplus$. Although the learning algorithms mentioned above can be used to infer the structure of such networks, some algorithms employ knowledge of these logical constraints in the learning process.

A widely used approach to learn a monotonic cancer progression network with a directed acyclic graph (DAG) structure and *conjunctive events* are *Conjunctive Bayesian Networks* (see CBNs, [23]). This model is a standard BN over Bernoulli random variables with the constraint that the probability of a node X taking the value 1 is zero if at least one of its parents has value 0. This defines a conjunctive relationship, in that all the parents of X must be 1 for X to possibly be 1. Thus, this model alone cannot represent noise, which is an essential part of any real data. In response to this shortcoming, *hidden CBNs* [24] were developed by augmenting the set of variables: to each CBN variable X, which captures the "true" state, is assigned a correspondence to a new variable Y that represents the observed state. Thus, each new variable Y takes the value of the corresponding variable X with a high probability, and the opposite value with a low probability. In this model, the variables X are latent, i.e., they are not present in the observed data, and have to be inferred from the observed values for the new variables. Learning is performed via a maximum likelihood approach and is separated into multiple iterations of two steps. First, the parameters for the current hypothesized structure are estimated using the *Expectation-Maximization* algorithm [25] and the likelihood given those parameters is computed. Second, the structure is perturbed using some *hill climbing* heuristic. In their work, the authors used the *Simulated Annealing* algorithm [26] for this step. These two steps are repeated until the score converges. However, the Expectation-Maximization algorithm only guarantees convergence to a likelihood local maximum and, thus, the overall procedure is not guaranteed to converge to the optimal structure.

Since CBNs represent the current benchmark for the reconstruction of cancer progression

[7]Specifically, the likelihood weight increases linearly, while the weight of the regularization term grows only logarithmically.

models from cross-sectional genomic data, their comparison with the approach we introduce is likely to be extremely informative (see §5).

3 A framework for prima facie causation

This section delves deeper into our framework for prima facie causation and its logical foundations. For the sake of clarity, we develop the presentation in steps of successively increasing complexity of the causal formulas: e.g., going from single-cause (i.e. "atomic") formulas, to conjunctive formulas consisting of atomic events to formulas in *Conjunctive Normal Forms* (CNF) (e.g., [('burning cigarette' $\wedge$ 'dried wood') $\vee$ ('lightning' $\wedge$ 'no rain') $\rhd$ 'forest fire'])[8]. The causal formulas are represented as a directed graph: $G = (V, E)$, where the nodes are the atomic events, and edges are between an event that appears positively as a literal in the formula describing the cause and an event that is its effect: $\forall_{c,e \in V}\ \langle c, e\rangle \in E$, iff c is a literal in φ and $\varphi \rhd e$.

Throughout this document, by *"real world"* we will refer to the concrete instance where data are gathered (as opposed to the counterfactual terminology of "possible worlds") and by *"topology"*, a combination of structural and quantitive probabilistic parameters.

3.1 Single-cause prima facie topologies

When at most a single incoming edge is assigned to each event (i.e., an event has at most one *unique cause* in the real world: $\forall_{e \in V} \exists!_{c \in V} c \rhd e$), we term this causal structure *single-cause prima facie topology*, a special and important case of the most general prima facie topology causal structures. Note that the general model can be represented as a direct acyclic graph (DAG) where each edge is a prima facie cause between a parent and its child. In the special case of the single-cause prima facie topology, the causal graphs are *trees* or, more generally, *forests* when there are disconnected components. Thus, each progression tree subsumes a distribution of observing a subset of the mutations in a cancer sample (see [27] for a detailed discussion).

In [27] the following propositions (summarized in Figure §2) were shown to hold for single-cause prima facie topologies, and used to derive an algorithm to infer tree (forests) models of cancer progression based upon the Definition §1 (by Suppes).

Statistical dependence. Whenever the PR holds between two events c and e, then the events are *statistically dependent* in a positive sense, i.e.

$$\mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c}) \iff \mathcal{P}(e \wedge c) > \mathcal{P}(e)\mathcal{P}(c)\,. \tag{3}$$

Mutuality. If c is a probability raiser for e, then so is the converse, i.e. $\mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c}) \iff \mathcal{P}(c \mid e) > \mathcal{P}(c \mid \overline{e})$

Natural ordering. For any two events c and e such that c is a probability raiser for e, a "natural" ordering arises to disentangle a causality relation, i.e.

$$\mathcal{P}(c) > \mathcal{P}(e) \iff \frac{\mathcal{P}(e \mid c)}{\mathcal{P}(e \mid \overline{c})} > \frac{\mathcal{P}(c \mid e)}{\mathcal{P}(c \mid \overline{e})}\,. \tag{4}$$

[8] The statement above may be shortened as 'burning cigarette' $\rhd$ 'forest fire.' The intended interpretation is that, 'burning cigarette' is an *insufficient but non-redundant part of an unnecessary but sufficient causal condition* (INUS) for 'forest fire,' as originally suggested by the philosopher J. Mackie.

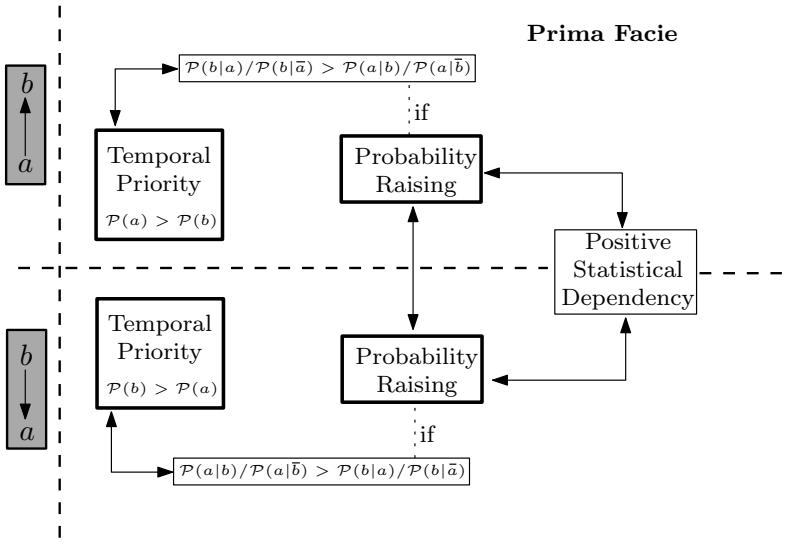


Figure 2: **Prima facie properties.** Properties of Suppes' definition of probabilistic causation (Definition §1) allow its rephrasing as: c is a *prima facie cause* of e *if the cause is a probability raiser of* e*, and it occurs more frequently.*

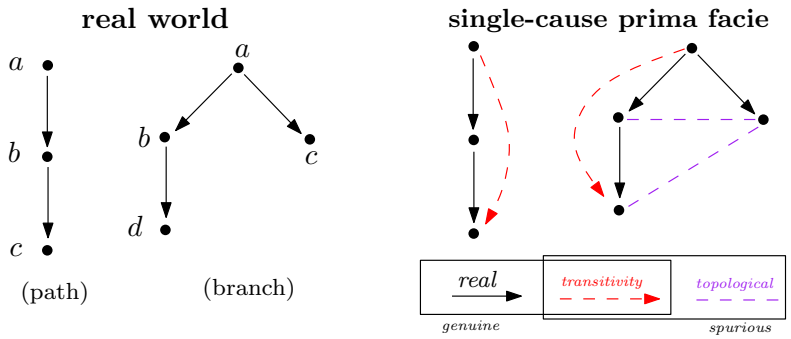
Putting together all these properties, it is natural to derive the following equivalent characterization of Definition §1: c is said to be a *prima facie cause* of e *if* c *is a probability raiser of* e*, and it occurs more frequently,* i.e.

$$c \rhd e \iff \mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c}) \quad \wedge \quad \mathcal{P}(c) > \mathcal{P}(e)\,. \tag{5}$$

Essentially, the assertion above restates that single-causes, involving only persistent events (see §1.4), lead to a model of real world time (t_c and t_e in Definition §1), which can be consistently *imputed* to the observed frequencies of events.

Consequent to this definition, we observe that (see our earlier discussion in Section §1.2) it is *necessary but not sufficient* to identify the causal real world processes (path or branch) and, thus, to solve causality *per se*. In fact, as it can be easily observed in the Figure §3, black arrows (consistently in the real world and in the topology) make this definition necessary, while red arrows (*spurious*, resulting from *transitivities*, because of the single-cause hypothesis) render the condition insufficient. We remark that red arrows will *always* be present to indicate potential *genuine* causes corresponding to real causes (which is the case when observations are statistically significant for the real world). Thus a correct inferential algorithm will have to select real causes among the potential genuine ones, a subset of prima facie causes.

A further discussion about spurious connections is now warranted. As discussed in §1.1, spurious causes may manifest through *spurious correlation* or *chance*. In the infinite sample size limit the "law of large numbers" eliminates the effect of chance; in other words, with large enough sample, chance by itself will not suffice to satisfy Definition §1. The former situation for spuriousness depends on the real world topology, and might appear under observation like a prima-facie/genuine cause in disguise, even with an infinite sample size (purple edges, for which the "temporal direction" has no causal interpretation, as it depends on the data and topology). For these reasons, a single-cause prima facie topology asymptotically will not contain *false negatives* (i.e. all real world causes are in the topology as Definition §1 is necessary) but



	temporal priority
path	$\mathcal{P}(a) > \mathcal{P}(b) > \mathcal{P}(c)$
branch	$\mathcal{P}(a) > \mathcal{P}(b) > \mathcal{P}(d)$ *and* $\mathcal{P}(a) > \mathcal{P}(c)$

Figure 3: **Single-cause prima facie topology.** Example of linear path and branching causal processes in the real world and corresponding single-cause prima facie topologies, according to Definition §1, with infinite sample size. We show all the genuine connections (red and black, directed by the temporal priority), and augment the topology with edges (purple, undirected) which might be suggested by the topology (or observations, if data were finite).

it might contain, depending on the real world topology, *false positives* (red or purple edges, as Definition §1 is not sufficient).

3.2 Conjunctive-cause prima facie inference

We denote by a Boolean conjunctive clause, a propositional formula composed of conjunctions of a set of literals: $\boldsymbol{c} = c_1 \wedge \cdots \wedge c_n$, which implies that n events $c_1, \ldots, c_n$ have occurred (in some unspecified order) so as to collectively cause some effect e (graphically pictured as in Figure §4), and we assume that each c_i $(1 \leq i \leq n)$ is an atomic event.

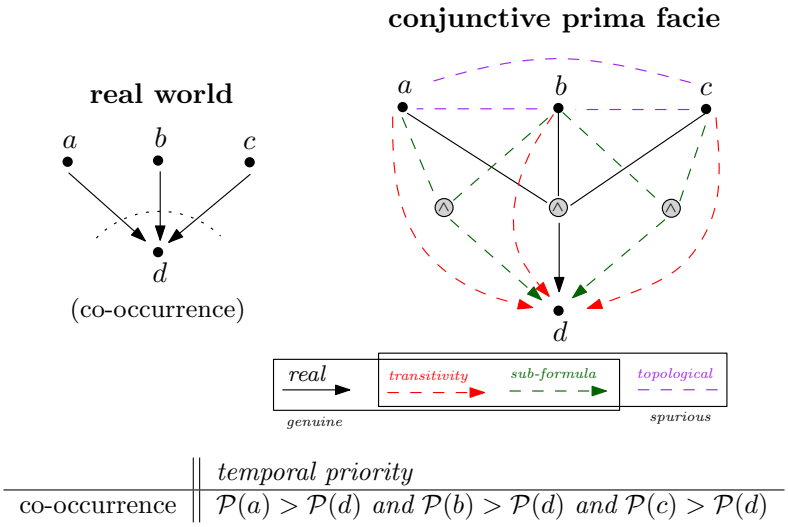
Suppes' notion of probabilistic causation (Definition §1) can be naturally extended to conjunctive clauses as in the following definition:

Definition 2 (Conjunctive probabilistic causation)**.** *For any conjunctive event* $\boldsymbol{c} = c_1 \wedge \ldots \wedge c_n$ *and* e*, occurring respectively at times* $\{t_{c_i} \mid i = 1, \ldots, n\}$ *and* t_e*, under the mild assumptions that* $0 < \mathcal{P}(c_i), \mathcal{P}(e) < 1$*, for any* i*, the conjunctive event* $\mathbf{c}$ *is a* prima facie conjunctive cause *of* e *(*$\mathbf{c} \rhd e$*) if all of its components* c_i *occur* before *the effect and their occurrences collectively* raises the probability *of the effect, i.e.*

$$\max\{t_{c_1}, \ldots, t_{c_n}\} < t_e \quad \textit{and} \quad \mathcal{P}(e \mid \boldsymbol{c}) > \mathcal{P}(e \mid \overline{\boldsymbol{c}})\,. \tag{6}$$

where $\mathcal{P}(e \mid \boldsymbol{c}) = \mathcal{P}(e \mid c_1 \wedge \cdots \wedge c_n)$ *and* $\mathcal{P}(e \mid \overline{\boldsymbol{c}}) = \mathcal{P}(e \mid \overline{c_1 \wedge \cdots \wedge c_n}) = \mathcal{P}(e \mid \overline{c}_1 \vee \cdots \vee \overline{c}_n)$.

This extension simply follows the semantics of conjunctive connectives, which states that *all causes* must occur *before* the effect, thus justifying the choice of picking the latest event, in time, prior to e to generalize Definition §1: namely, the $\max\{\cdot\}$ operation applied to the causal events. Clearly, this definition retains the semantics of single-cause prima facie unchanged, as it is just a special case with $\boldsymbol{c} = c$ and $\max\{t_{c_i}\} = t_c$. Unfortunately, as before, it still has the same weakness that it is *necessary but not sufficient* to identify conjunctive-causal relations, and hence lacks the power to define causality *per se*.



	temporal priority
co-occurrence	$\mathcal{P}(a) > \mathcal{P}(d)$ *and* $\mathcal{P}(b) > \mathcal{P}(d)$ *and* $\mathcal{P}(c) > \mathcal{P}(d)$

Figure 4: **Conjunctive-cause prima facie topology.** Example of conjunctive real world process (a *and* b *and* c cause d). We show the conjunctive-cause prima facie topology according to Definition §2, with all genuine connections and infinite sample size. The topology is augmented by logical connectives, as done for Figure §3.

The properties of single-causes prima facie topologies extend appropriately to conjunctive topologies – a fact proven in the Proof Section at the end of this document, along with all the other properties and theorems that appear in this Supplementary Materials.

Proposition 1. *The properties of* statistical dependence, mutuality *and* natural ordering *for single-causes are still valid for conjunctive clauses.*

In this case some caution must be exercised in distinguishing between prima facie single or conjunctive causes. As shown in Figure §4, in fact, for a simple conjunctive clause in the real world (a *and* b *and* c) the following conjunctive clauses

$$a \wedge b \rhd d \qquad\qquad a \wedge c \rhd d \qquad\qquad b \wedge c \rhd d$$

as well as the single causes $a \rhd d$, $b \rhd d$ and $c \rhd d$, are prima facie. The single causes can be *spurious* or *transitive*, as in Figure §3. But now, we will also call *spurious sub-formulas* the conjunctive clauses that are *syntactically strictly sub-formulas* of $a \wedge b \wedge c \rhd d$, i.e., the only formula we would like to infer. Notice that as in branch processes, topology-dependent spurious causes might appear because of spurious correlations; in the Figure §3, we have not shown other potential spurious causes, as what we depict is just a one-level conjunctive network. These causal relations could include general *spurious formulas* constituting of a sub-formula and any of its parents. Similarly, spurious causes due to chance will vanish asymptotically as sample size grows to infinity. Summarizing, we note that a conjunctive topology, just as in the single-cause framework, will not contain false negatives (i.e., all real world causes will be in the topology) but it might contain, depending on the real world topology, false positives (red, green or purple edges).

Before concluding, we note that the total number of potential formulas and transitivities is

exponential in the size of $|G| = n$, that is

$$\sum_{i=1}^{n-1} \binom{n-1}{i} = 2^{n-1} - 1\,.$$

Notice that this is a lower bound accounting only for the level of the connective, and is expected to grow further when more complex real world processes are considered. Finally, as shown in Figure §3, the number of spurious causes due to topology (purple edges), are quadratic in the formula size, being

$$2\binom{n-1}{2} = (n-1)(n-2)\,.$$

This complexity hints at the fact that an exhaustive search of all the possible conjunctive formula is not feasible, in general.

3.3 Generalization to formulas in conjunctive normal form

Next, consider a formula in *conjunctive normal form* (CNF)

$$\varphi = \boldsymbol{c}_1 \wedge \ldots \wedge \boldsymbol{c}_n,$$

where each $\boldsymbol{c}_i$ is a *disjunctive clause* $\boldsymbol{c}_i = c_{i,1} \vee \ldots \vee c_{i,k}$ over a set of literals, each literal representing an event (a Boolean variable) or its negation. By following the same approach as used earlier to extend Suppes' Definition §1 from single to conjunctive clauses, we define $\varphi \rhd e$.

Definition 3 (CNF probabilistic causation)**.** *For any CNF formula φ and e, occurring respectively at times t_φ and t_e, under the mild assumptions that $0 < \mathcal{P}(\varphi), \mathcal{P}(e) < 1$, φ is a* prima facie cause *of e if*

$$t_\varphi < t_e \quad \textit{and} \quad \mathcal{P}(e \mid \varphi) > \mathcal{P}(e \mid \overline{\varphi})\,. \tag{7}$$

As before, this definition subsumes Definition §2 and is thus *necessary but not sufficient* to identify causal relations, hence lacking the power to solve causality *per se*.

Clearly, in this case, the number of prima facie (including both genuine and spurious) causes grows combinatorially much more rapidly than the simplest case of a unique conjunctive clause (Section §3.2); this situation is rather alarming, since even the simplest case already produces an exponentially large set of prima facies causes in terms of the number of events. In this case, in fact, further causal relations emerge as a result of mixing events from all the clauses of φ. CNF formulas follow analogous properties as single and conjunctive topologies, as shown below.

Proposition 2. *The properties of* statistical dependence*,* mutuality *and* natural ordering *for single and conjunctive prima facie topologies extend to CNF formulas* mutatis mutandis*.*

We conclude this section with two final comments about CNF formulas, their relation with background contexts, and the notion of timing in Definition §3.

Our first comment concerns Cartwright's idea of background contexts as a conjunction of independent factors (Section §1). For illustrative purposes, consider the formula $(a \wedge b) \vee c \rhd d$, which is in *disjunctive normal form* (DNF). If, for example, we were to evaluate the claim $a \rhd d$, the (unique) background context would be the atomic event c, being b-dependent when a causes d. A symmetric situation holds, were we to evaluate $b \rhd d$. In light of this discussion note that, if we convert the formula to its CNF analogue $(a \vee c) \wedge (b \vee c) \rhd d$, we need to correctly interpret the roles of sub-formulas $a \vee c$ and $b \vee c$ in identifying a background context, c. It follows immediately

that, for any CNF formula, the atomic events of all the disjunctive clauses in the equivalent DNF formula provide all the possible background contexts à-la-Cartwright.

Our second comment concerns timing in the real world. Consider the CNF formula above, denote it as φ and recall that Definition §3 requires $t_\varphi < t_d$. One might wonder whether a trivial time-ordering relation exists, whose complexity is linear with respect to all the operators in φ. Were it so, we would be able to parse φ into its constituents, and recursively express the temporal relations as a direct function of those relations that hold for its sub-formulas. Unfortunately, this appears not to be the case, except when the underlying syntax is restricted to certain specific operators (e.g., conjunctions). Thus appropriate care must be taken in implementing a model of real world time. Thus, an algorithm, working on the illustrative example of the previous paragraph, cannot conclude any ordering about $t_{a\vee c}$, $t_{b\vee c}$ and t_d, solely by looking at the observed probabilities of their atomic events – instead it must gather the correct information for certain sub-formulas at the level of their connective (the $\vee$ in this case). A general rule that avoids these difficulties and devises a correct and efficient timing-inference algorithms, may be stated as follows: it is *always safe* to model probabilistic causation in terms of whole formulas, while permitting *compositional* reasoning over sub-formulas, only when the syntax is restricted to certain Boolean connectives. Further related comments appear in the next sections, where we describe the complete algorithm.

4 An inference algorithm

The structure of the reconstruction problem is as follows. Assume that we have a set G of n mutations (*events*, in probabilistic terminology) and m samples, represented as a cross-sectional dataset, i.e., without explicit timing information, in an $m \times n$ binary matrix $D \in \{0,1\}^{m\times n}$ in which an entry $D_{k,l} = 1$ if the mutation l was observed in sample k, and 0 otherwise. Note that dataset lacking explicit timing information are typical: for instance, in cancer patient data. However, we work in the same setting as that used in [28, 29, 23, 27] already.

To introduce the algorithm, few more additional notations are required: we denote by $\mathcal{U}$ the *universe* of all possible causal claims $\varphi \rhd e$, where φ is a CNF formula over the events in D (thus $G \subseteq \mathcal{U}$) and e is an atomic event. With $\mathcal{C} \subset \mathcal{U}$ we denote all the causal claims whose formulas are conjunctive over atomic events, that is they do not contain disjunctions. For a general CNF formula φ we denote by $\text{chunks}\,(\varphi)$ its set of disjunctive clauses. For example, $a \wedge b \rhd e \in \mathcal{C}$ while $(a \vee \bar{b}) \wedge (c \vee d) \wedge e \rhd f \notin \mathcal{C}$ and $\text{chunks}\left((a \vee \bar{b}) \wedge (c \vee d) \wedge e\right) = \{(a \vee \bar{b}), (c \vee d), e\}$.

Inferred structures. Our algorithm reconstructs a general DAG from the input data. Not too surprisingly, it shares many structural and algorithmic properties with the Conjunctive Bayesian Networks approach of [23] – especially in the context of cancer progression models. However, our algorithm faces no obstacle in spontaneously inferring from the input data various sub-structures of a DAG, e.g., forests – or, more specifically, trees – although it has no "hard-coded" policies for doing so. Thus, we expect the algorithm to be applicable in a context-agnostic manner and compete well with other approaches, which are not *a priori* restricted from having advantageous structural information, e.g., [27, 28, 30, 29].

In contrast to [23], our DAGs can build on arbitrary CNF formulas, using the strategy that *disjunctive clauses* are first summarized by unique DAG nodes. As an example, a formula $(a\vee b)\wedge c\wedge d$ will be modeled with three nodes: one for $(a\vee b)$, the aggregated disjunction, one for c and one for d. The reasons we do not explicitly handle disjunctions are discussed subsequently.

In the following, we will denote a *progression DAG* as $\mathcal{D} = (N, \pi)$ where $N \subseteq \mathcal{U}$ is the set of nodes (e.g, *mutations* or formulas) and $\pi : N \to \wp(N)$ is a function associating to each node j

its *parents* $\pi(j)$. This model yields the following.

Definition 4 (DAG causal claims)**.** *A* $\mathcal{D} = (N, \pi)$ *models the* causal claims

$$\bigcup_{j \in N} \Big\{ (c_1 \wedge \ldots \wedge c_n) \rhd j \mid \pi(j) = \{c_1, \ldots, c_n\} \Big\},$$

where $c_1 \wedge \ldots \wedge c_n$ *is a CNF formula and any* c_j *is either a ground event or a disjunction of events.*

Going back to the example above, in our DAG we would have $\pi(j) = \{(a \vee b), c, d\}$ whose underlying causal claim would be $(a \vee b) \wedge c \wedge d \rhd j$.

Each DAG is augmented with a labeling function $\alpha : N \to [0, 1]$ such that $\alpha(i)$ is the *independent probability* of observing mutation i in a sample, whenever *all of its parent* mutations are observed (if any). Each DAG induces a *distribution* of observing a subset of events in a set of samples (i.e., a probability of observing a certain *mutational profile* in the context of our application), as defined below.

Definition 5 (DAG-induced distribution)**.** *Let* $\mathcal{D}$ *be a DAG and* $\alpha : N \to [0, 1]$ *a labeling function,* $\mathcal{D}$ *generates a distribution where the probability of observing* $N^* \subseteq N$ *events is*

$$\mathcal{P}(N^*) = \prod_{x \in N^*} \alpha(x) \cdot \prod_{y \in N \setminus N^*} \Big[1 - \alpha(y) \Big] \tag{8}$$

whenever $x \in N^*$, $\pi(x) \subset N^*$, *and* 0 *otherwise.*

Notice that this definition, as expected, is equivalent to the one used in [23] and retains a tree-induced distribution such as those used in [27, 28, 30]. Further, notice that a sample which contains an event but not all of its parents has a zero probability, thus subsuming the conjunctive interpretation of DAGs. These kinds of samples, which represent "irregularities" with respect to $\mathcal{D}$, might be generated when adding false positives/negatives to the sampling strategy. Finally, the fact that we allow nodes to be disjunctive formulas, extends this DAG definition to express causal claims with generic CNF formulas.

Inference confidence: bootstrap and statistical testing. We provide a statistical foundation to our inferences, which employ such classical techniques as: *bootstrap* [31, 32], and the *Mann-Whitney U test* [33].

In data preprocessing we use *bootstrap with rejection resampling*; according to §1.4, we proceed as follows to estimate a distribution of the marginal and joint probabilities, for each event: (i) we sample with repetitions rows from the input matrix D (bootstrapped dataset), (ii) we next estimate the distributions from the observed probabilities, and finally, (iii) we reject values which do not satisfy $0 < \mathcal{P}(i) < 1$ and $\mathcal{P}(i \mid j) < 1 \vee \mathcal{P}(j \mid i) < 1$, and iterate restarting from (i). We stop when we have, for each distribution, at least 100 values.

Any inequality (i.e., checking temporal priority and probability raising) is estimated as follows: We perform the Mann-Whitney U test with p-values set to 0.05. This is a non-parametric test of the null hypothesis that two populations are the same against an alternative hypothesis, and is especially useful to understand wether a particular population, e.g., $\mathcal{P}(i)$, tends to assume larger values than the other, e.g., $\mathcal{P}(j)$. By employing this test, which need not assume Gaussian distributions for the populations, confidence p-values for both temporal priority and probability raising are computed.

Once a DAG model is inferred with the algorithm described in the next section, both *parametric and non-parametric bootstrapping methods* can be used to assign a confidence level to its

respective claims, and ultimately, to the overall causal model. Essentially, these tests consist of using the reconstructed model (in the parametric case), or the probabilities observed in the dataset (in the non-parametric case) to generate new synthetic datasets, which are then reused for reconstructin of the progressions (see, e.g., [32] for an overview of these methods). The confidence is given by the number of times the DAG or any of its claim is reconstructed from the generated data.

4.1 CAPRI: a hybrid algorithm for general CNF formulas

Building upon the framework presented earlier in §3, we present here a novel algorithm to infer cancer progression models from cross-sectional data. The algorithm is hybrid in the sense that it combines a structure-based (as of Definition §3) approach with a likelihood-fit constraint and, according to its input, infers causal claims with various logical expressivity. Its computational complexity, which is highly dependent on the expressivity of the claims, as well as its correctness are discussed in the next section.

CAncer PRogression Inference (CAPRI, Algorithm §1) requires as its input, a matrix D and, optionally, a set of k input causal claims $\Phi = \{\varphi_1 \rhd e_1, \ldots, \varphi_k \rhd e_k\}$, where each φ_i is a CNF formula and $\varphi_i \not\sqsubseteq e_i$. Here $\sqsubseteq$ represents the usual *syntactical* ordering relation among atomic events and formulas, e.g., $a \sqsubseteq (a \vee b) \wedge c \wedge d$, and is simply required to disallow malformed input claims, which would vacuously be labeled as prima facie causality (as of Definition §3) but would have no real causal meaning, i.e., in the example above it makes no sense to say that "a causes $(a \vee b) \wedge c \wedge d$." The augmented input Φ, which contains claims of the most complex type CAPRI can infer, is optional in the sense that, if $\Phi = \emptyset$, the algorithm is able to infer "all" conjunctive causal claims over atomic events (e.g., claims $a \wedge b \wedge c \rhd e$ in $\mathcal{C}$), but not general CNF ones.

CAPRI starts by performing a *lifting operation*[9] over D, and then build a DAG $\mathcal{D}$. Lifting operation evaluates each CNF formula φ_i for all input causal claims in Φ and its result, a lifted D, is an extended input matrix for the algorithm. As an example consider a claim $\Phi = \{(a \vee \bar{b}) \wedge (c \vee d) \wedge e \rhd f\}$, the result of lifting for an input matrix D over $a, \ldots, f$ is

$$D = \left[\begin{array}{cccccc} a & b & c & d & e & f \\ \hline 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \end{array}\right], \qquad D(\Phi) = \left[\begin{array}{cccccc|c} a & b & c & d & e & f & \varphi \\ \hline 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{array}\right],$$

since $\varphi = (a \vee \bar{b}) \wedge (c \vee d) \wedge e$ and, e.g., $(1 \vee \overline{0}) \wedge (1 \vee 0) \wedge 0 \equiv 0$. After the lifting, $\mathcal{D}$ is built by individually including in its set of nodes all the disjunctive sub-formulas of such CNF formulas, plus G. In the preceding example, $\{(a \vee \bar{b}), (c \vee d), e\}$ are nodes in $\mathcal{D}$ (note that $e \in G$). Notice that $D(\Phi) = D$ and $N = G$ if $\Phi = \emptyset$.

Subsequently, the parent function (i.e., the edges in $\mathcal{D}$) is built by pair-wise implementation of Definition §3, which has been shown to subsume also Definitions §1 and §2. For the sake of simpler exposition, we make use of the coefficients $\Gamma_{i,j}$ and $\Lambda_{i,j}$ to evaluate temporal priority and probability raising, respectively, which are required to be strictly positive by Definition §3. We distinguish two cases: (i) when we are evaluating a causal claim directly involving an atomic

[9] The term lifting was originally introduced to denote the lifting among partial and complete orderings in denotational semantics. Later on, it was used by Pearl [11] to denote the removal of an equation $Y = f(\cdot)$ with a fixed association $Y = y$, i.e., a solution for $f(\cdot)$ which reduces the space of solutions of the overall problem. Here we are close to Pearl in the sense that we lift D to $D(\Phi)$ by considering the claims in Φ, and not all possible causal claims.

event, or (ii) a chunk of an input formula. When a claim "i causes j" is evaluated and $i \in G$, we just require that Definition §3 is satisfied; if so i is a prima facie cause of j and we add it to $\pi(j)$. When we do the same for an input formula φ, if it is prima facie for an event j we add φ via all its constituting chunks to $\pi(j)$. This is required by the fact that the DAG $\mathcal{D}$ is built by chunking input formulas, while the lifting operation is performed on whole formulas; in reference to the examples above, when φ is prima facie to f, we would add $(a \vee \bar{b})$, $(c \vee d)$ and e to $\pi(f)$. Finally, since we are interested in claims with the rightmost part an atomic event, we force $\pi(j) = \emptyset$ for any $j \notin G$. In case of the preceding input, for instance, we would not consider any incoming edge in $(a \vee \bar{b})$ and $(c \vee d)$, while we would consider edges incoming in e solely from an atomic event. As for labeling, note that no label is assigned to this kind of nodes. Finally, since this construction is consistent with our approach and the conjunctive interpretation of $\mathcal{D}$, once the steps defined in equations (§9– §10) have been performed, $\mathcal{D}$ is indeed a *prima facie DAG*.

As prima facie causality provides only a necessary condition, we must attempt filtering out all *spurious causes* that might have been included in $\mathcal{D}$. The underlying intuition is as follows: for any prima facie structure, spurious claims will contribute to reduce the likelihood-fit relative to true claims, and thus a standard maximum-likelihood fit can be used to select and prune the prima facie DAG. Based on all the discussion made in §2.2.2, it should be clear that a *regularization term* is necessary to avoid overfitting. In fact, if simple log-likelihood were used, we should expect that the best model is actually the prima facie structure. For this reason, we adopt the regularization score discussed in §2.2.2, namely *Bayesian Information Criterion* (BIC), which implements *Occam's razor* by combining log-likelihood fit with a *penalty criterion* proportional to the log of the DAG size via *Schwarz Information Criterion* [34].

Note that with $\Phi = \emptyset$ only conjunctive causal claims in $\mathcal{C}$ are inferred by our algorithm, since the set of nodes of $\mathcal{D}$ is $N = G$. Analysis of complexity, correctness and expressivity of CAPRI can now be presented.

4.2 Complexity, correctness and expressivity of CAPRI

Complexity. The previous sections have stressed the rapidity with which the set of causal claims (or formulas) grow for a given model, thus making their inference highly intractable. However, this complexity is intrinsic to the problem; or put alternatively, it is independent of the underlying theory of causation. Unlike the heuristic approaches, commonly used by many others to infer general causal claims, we adopt a twofold approach. To infer simple claims (i.e., single or conjunctive causes, at most), CAPRI's execution is self-contained (i.e., no input besides D is required) and polynomial in the size of D. Instead, we limit the number of inferable general causal claims (i.e., CNF), by requiring that they be specified as an input to the algorithm in Φ; in this case CAPRI tests, with a polynomial cost, those claims plus the simple ones, and its complexity spans over many orders of magnitude according to the structural complexity of the input set Φ, as further elaborated in the following theorem.

Theorem 1 (Asymptotic complexity)**.** *Let $|G| = n$ and $D \in \{0,1\}^{m \times n}$ where $m \gg n$, and let N the nodes in the DAG returned by CAPRI, the* worst case *time and space complexity of building a prima facie topology is, ignoring the cost of bootstrap:*

[10] Although CAPRI is equipped with bootstrap testing it is still possible to encounter various degenerate situations. In particular, for some pair of events it could be that temporal priority cannot be satisfactorily resolved, i.e. there is no significant p-value for any edge orientation. Thus, loops might be present in the inferred prima facie topology. Nonetheless, some of these could be still disentangled by PR, while some might remain, albeit rarely. To remove such edges we suggest to proceed as follows: (*i*) sort these edges according to their p-value (considering both temporal priority and probability raising), (*ii*) scan the sorted list in decreasing order of confidence, (*iii*) remove an edge if it forms a loop.

Algorithm 1 *CAncer PRogression Inference* (CAPRI)

1: **Input:** A set of events $G = \{g_1, \ldots, g_n\}$, an $m \times n$ matrix $D \in \{0,1\}^{m \times n}$ and k CNF causal claims $\Phi = \{\varphi_1 \rhd e_1, \ldots, \varphi_k \rhd e_k\}$ where, for any i, $e_i \not\sqsubseteq \varphi_i$ and $e_i \in G$;

2: [*Lifting*] Define the *lifting of D to $D(\Phi)$* as the augmented matrix

$$D(\Phi) = \left| \begin{array}{cccccc} D_{1,1} & \ldots & D_{1,n} & \varphi_1(D_{1,\cdot}) & \ldots & \varphi_k(D_{1,\cdot}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ D_{m,1} & \ldots & D_{m,n} & \varphi_1(D_{m,\cdot}) & \ldots & \varphi_k(D_{m,\cdot}) \end{array} \right|. \tag{9}$$

by adding a column for each $\varphi_i \rhd c_i \in \Phi$, with φ_i evaluated row-by-row, define the coefficients

$$\Gamma_{i,j} = \mathcal{P}(i) - \mathcal{P}(j), \qquad \text{and} \qquad \Lambda_{i,j} = \mathcal{P}(j \mid i) - \mathcal{P}(j \mid \bar{i}), \tag{10}$$

pair-wise over $D(\Phi)$;

3: [*DAG structure*] Define a DAG $\mathcal{D} = (N, \pi)$ where[10]

$$N = G \cup \left(\bigcup_{\varphi_i} \text{chunks}\,(\varphi_i) \right), \qquad \pi(j \notin G) = \emptyset;$$
$$\pi(j \in G) = \Big\{ i \in G \mid \Gamma_{i,j} \wedge \Lambda_{i,j} > 0 \Big\} \cup \Big\{ \text{chunks}\,(\varphi) \mid \Gamma_{\varphi,j} \wedge \Lambda_{\varphi,j} > 0,\ \varphi \rhd j \in \Phi \Big\}. \tag{11}$$

4: [*DAG labeling*] Define the labeling α as follows

$$\alpha(j) = \begin{cases} \mathcal{P}(j), & \text{if } \pi(j) = \emptyset \text{ and } j \in G; \\ \mathcal{P}(j \mid i_1 \wedge \ldots \wedge i_n), & \text{if } \pi(j) = \{i_1, \ldots, i_n\}. \end{cases}$$

5: [*Likelihood fit*] Filter out all spurious causes from $\mathcal{D}$ by likelihood fit with the regularization BIC score and set $\alpha(j) = 0$ for each removed connection.

6: **Output:** the DAG $\mathcal{D}$ and α;

- $\Theta(mn)$ *time and* $\Theta(n^2)$ *space, if* $\Phi = \emptyset$*;*
- $\Theta(|\Phi|mn)$ *time and* $\Theta(|\Phi|m)$ *space, if* $\Phi \subset \mathcal{U}$ *and* $|N| \ll m$ *(i.e., there are sufficiently many samples to characterize the input formulas);*
- $\mathcal{O}(2^{2^n})$ *time and space, if* $\Phi = \mathcal{U}$*.*

Thus, the overall complexity of CAPRI is any of the above, plus the cost of likelihood fit.

As shown above, the algorithmic complexity spans over many orders of magnitude according to the structural complexity of the input set Φ which determines the number of nodes in the returned DAG, i.e. $|N|$. Hence, aside from the cost of likelihood fit, the cost of the algorithm is polynomial only if Φ is polynomial in the number of input samples and atomic events. This observation forewarns one of the hazard of a brute force approach, attempting to test all possible causal claims. Generally speaking, despite the price of possibly "missing" some real causal claims, one should be able to identify most relevant causal structures by exploiting domain-knowledge, biological priors, and empirical/statistical estimations in selecting reasonable input Φ (e.g., focusing on certain key driver-mutations over the others). Note that this problem's inherent computational intractability does not negate the power of the algorithmic automation, as proposed here, relative to what is currently achievable with manual analysis.

Correctness and expressivity. Let $\mathcal{W} \subseteq \mathcal{U}$ be the set of true causal claims in the real world, which we seek to infer (in the tests of our algorithm on synthetic data, $\mathcal{W}$ will be known, once a DAG to generate its input data is fixed). Here, we investigate the relation between $\mathcal{W}$ and the set of causal claims retrieved by our algorithm, as a function of sample size m and the presence of false positives/negatives which are assumed to occur at rates ϵ_+ and ϵ_- (we discuss in the Results section how the sampling of the input data is affected by such rates).

Below Σ denotes the set of causal relations, implicit in the DAG $\hat{\mathcal{D}}$ returned by our algorithm for an input set Φ and a matrix D; we write this fact as $D(\Phi) \Vdash \Sigma$. Such claims are evaluated as in Definition §4. We prove the following.

Theorem 2 (Soundness and completeness)**.** *When the sample size $m \to \infty$ and the data is uniformly affected by false positives and negatives rates $\epsilon_- = \epsilon_+ \in [0, 1)$, if the input given is a superset of the true causal claims, then CAPRI reconstructs exactly the true causal formulas $\mathcal{W}$, that is, if $\mathcal{W} \subset \Phi$ then $D(\Phi) \Vdash \mathcal{W} \cap \Phi$.*

Notice that if it could be assumed that Φ characterizes $\mathcal{W}$ well, then all real causal claims are in Φ, and the corollaries below follows immediately.

Corollary 1 (Exhaustivity)**.** *Under the hypothesis of the above theorem $D(\mathcal{U}) \Vdash \mathcal{W}$.*

Corollary 2 (Least Fixed Point)**.** *$\mathcal{W}$ is the* lfp *of the monotonic transformation*

$$\bigsqcup_{\Phi} D\Big(\Phi\Big) \equiv D\Big(\bigsqcup_{\Phi} \Phi\Big) \Vdash \mathcal{W}\,.$$

Since a direct application of this theorem incurs a prohibitive computational cost, it only serves to idealize the ultimate power of the framework we have proposed. That is, the theorem only states that CAPRI is able to select only the true causal claims asymptotically, as the size of $\mathcal{U}$ grows, albeit exponentially. It also clarifies that the algorithm is able to "filter out" all the spurious causal claims (true negatives), and produces the true positives from the set of the genuine causal claims more and more reliably as a function of the computational and data resources.

Now we restrict our attention to conjunctive clauses in $\mathcal{C}$ – i.e., those formulas which are defined only on atomic events – so as to enable a fair comparison with [23].

Theorem 3 (Inference of conjunctive clauses)**.** *Let $\Phi = \emptyset$; as before, when the sample size $m \to \infty$ and the data is uniformly affected by false positives and negatives rates $\epsilon_- = \epsilon_+ \in [0, 1)$, then only conjunctive clauses on atomic events are inferred, which are either true or spurious for general CNF formulas. That is: if $D(\emptyset) \Vdash \Sigma$ then $\Sigma \subseteq \mathcal{C}$. Furthermore,*

1. *$\Sigma \cap \mathcal{W}$ are true claims and*
2. *for any other claim $\alpha \rhd e \in (\Sigma \setminus \Sigma \cap \mathcal{W})$ there exist $\beta \rhd e \in \mathcal{W} \setminus \mathcal{C}$ such that β screens off α from e.*

This theorem states that even if one is neither willing to pay the cost of augmenting the input set of formulas nor is one able to find suitable formula to augment, the algorithm is still capable of inferring conjunctive clauses, whose members are either genuine or a conjunctive sub-formula of a more complex genuine CNF formula β (regardless of whether a cause of the second kind is considered to be spurious).

An immediate corollary of these two theorems is that the algorithm works correctly, when it is fed with all possible conjunctive formulas.

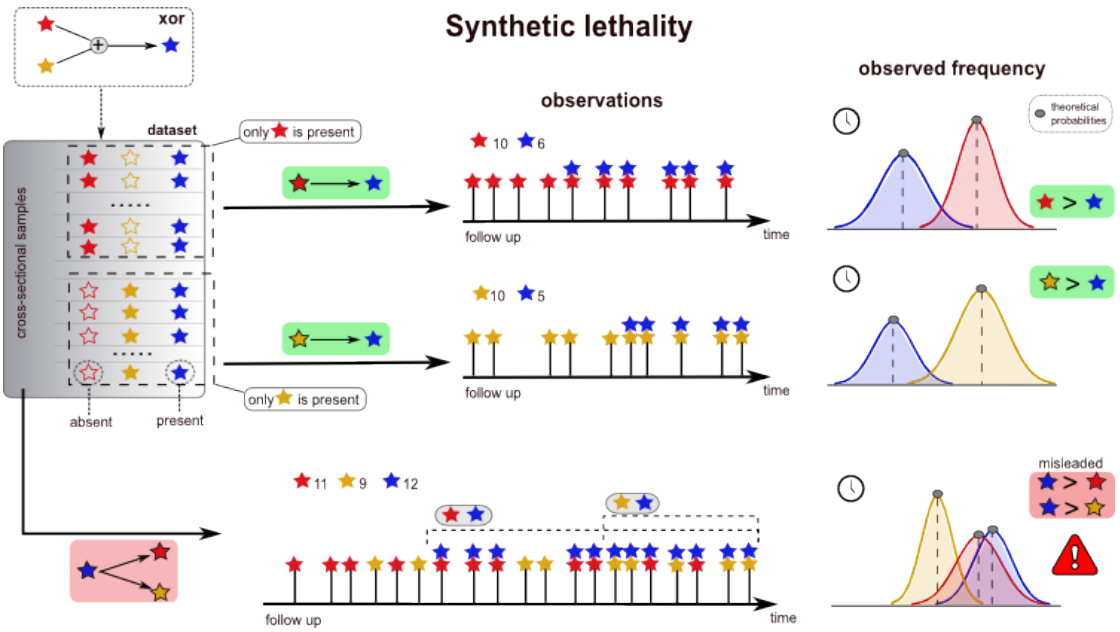


Figure 5: **Caveats in inferring synthetic lethality relations.** For a synthetic lethality causal relation among a and b towards c if one considers a dataset of aggregated samples the risk of misleading the *temporal priority* relation among a, b and c is high. If one were to know, a priori, that $a \oplus b$ is part of the claim, one could separate data and work safely. Unfortunately, being unknown a priori, one only relies on domain knowledge, biological priors or hypothesis testing.

Corollary 3. *Under the hypothesis of the above theorems,* $D(\emptyset) \Vdash \Sigma \iff D(\mathcal{C}) \Vdash \Sigma$.

In practice, though still exponential, this algorithm is certainly less computationally intensive, when using $\mathcal{C}$ than with $\mathcal{U}$, as it can trade off computational complexity against expressivity of the inferred causal claims.

One final comment is due at this point. In the context of automatic inference of logical formulas *expressivity* of the inferred claims relates to *compositional inference*. In particular, it is easy to see that for a disjunctive formula $\boldsymbol{c}_1 \vee \ldots \vee \boldsymbol{c}_n$, the following holds

$$\boldsymbol{c}_1 \vee \ldots \vee \boldsymbol{c}_n \rhd e \quad \not\Rightarrow \quad \forall_{\boldsymbol{c}_i} \boldsymbol{c}_i \rhd e,$$

which is the reason why we cannot compositionally infer full CNF formulas by reasoning over their constituents (i.e., any $\boldsymbol{c}_i$ might not satisfy the prima facie definition on its own). Thus, we have to rely on the hypothesis set Φ, unless one could assume to know *a priori* the formulas and hence the background contexts (i.e., any other $\boldsymbol{c}_j$, for $j \neq i$), which poses a *circularity* issue. An instance of this constraint is of particular importance with respect to cancer: for example, in modeling *synthetic lethality* (see Figure §5) which can be expressed as $c_1 \oplus c_2 \rhd e$ where $c_1 \oplus c_2 = (c_1 \wedge \bar{c}_2) \vee (\bar{c}_1 \wedge c_2)$.

Further commentary and comparison with the literature. The algorithm, defined here, can be applied to infer tree or forest models of progression, and can be evaluated empirically against other approaches in the literature which are specifically tailored for tree/forests [27, 28,

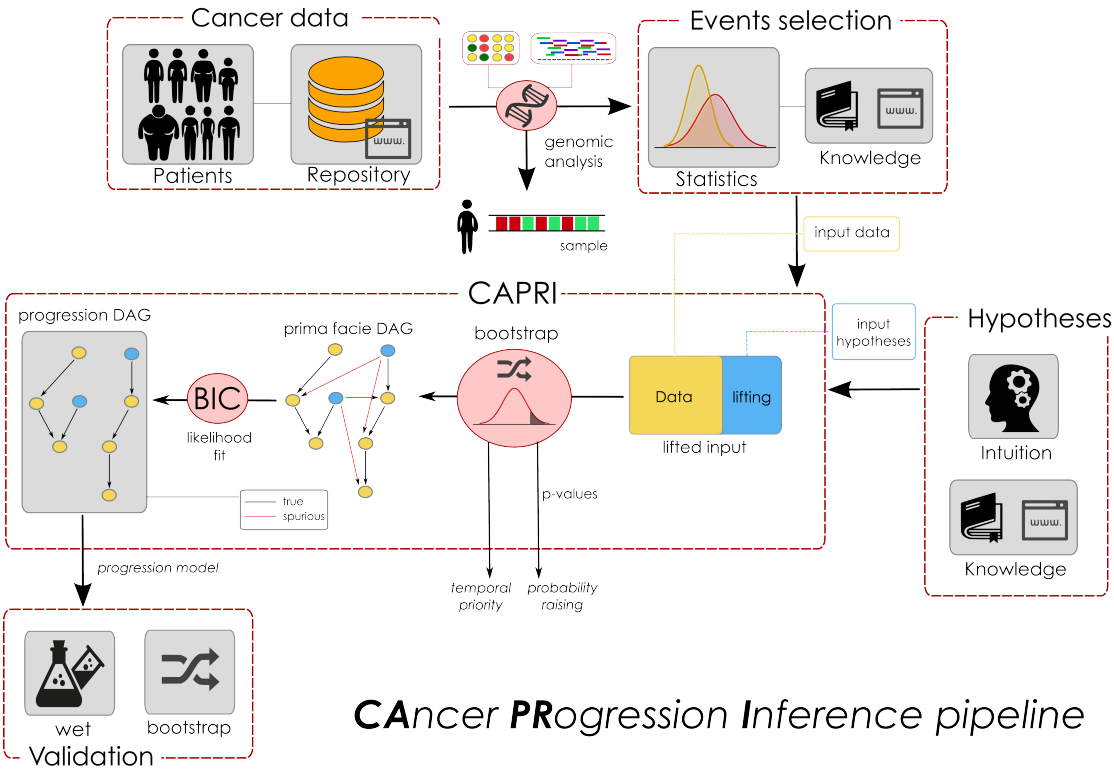


Figure 6: **Pipeline for CAPRI.** The pipeline starts with data gathering, either experimentally or via shared repositories, and genomic analysis to create, e.g., somatic mutation or Copy-Number Variations profiles for each sample. Then, events must be selected via statistical analysis and biological priors, to construct a suitable input data matrix D which satisfies CAPRI's assumptions. Hypothesis of any causal claim can then be generated, based on prior knowledge. CAPRI is then executed, which results in p-values for temporal priority and probability raising to be returned, along with the inferred progression model. Validation concludes the pipeline.

30]. All these approaches have the same quadratic complexity (in the number of events in $|G|$) and, just as with our CAPRI, have been shown to converge asymptotically to the correct tree, even in the presence of noisy observations. Despite asymptotic equivalence, the algorithms differ in performance under various settings of finite data (usually, synthetic), as reported extensively in our earlier publication [27]. The simpler algorithm, *CAncer PRogression Extraction with Single Edges* (CAPRESE, [27]), differs from CAPRI, as it relies on a score based on probability raising with a *shrinkage estimator*, which intuitively corrects for the sample size and noise (see [31] and [32]). By comparing the current algorithm with the one in [27], we directly shed light on the complexity and expressivity trade-offs between two very related algorithms; see next section.

5 Results: synthetic data

We next describe the details of the (a) setting, in which various empirical comparisons were carried out (reported in the next section), (b) generative models for synthetic data and finally,

(c) performance metrics used for comparison. A general pipeline for CAPRI's usage is depicted in Figure §6. CAPRI is implemented in the open source R package `TRONCO` (second version, available at standard R repositories).

Setting for comparison. The performance of all the algorithms were assessed with four different types of topologies: (i) *trees*, (ii) *forests*, (iii) *DAGs without disconnected components* and (iv) *DAGs with disconnected components*. Irrespective of the topology considered, we exclusively used atomic events, which implies that the kind of causal claims, we could experiment with, are either single or conjunctive. Based on Corollary §3, it sufficed to run CAPRI with $\Phi = \emptyset$. This is consistent with the fact that our algorithm can infer more general formulas if an input "set of putative causes, $\Phi \neq \emptyset$" is given in addition – a fact which could have biased our analysis in our favor in the more general situation. For the sake of completeness, however, we also tested specific CNF formulas, as shown in the next sections.

Type $(i - ii)$ topologies are DAGs constrained to have nodes with a unique parent; condition (i) further restricts such DAGs to have no disconnected components, meaning that all nodes are reachable from a starting root r. Practically, condition (i) satisfies $|\pi(j)| = 1$ for $j \neq r$, and $\pi(r) = \emptyset$, while in (ii) we allow more roots to be present. This kind of topologies can be either reconstructed with ad-hoc algorithms [27, 28, 30] or general DAG-inference techniques [21, 20, 35, 34, 23]. Type $(iii - iv)$ topologies are DAGs which have either a unique starting node r, or a set of independent sub-DAGs. Similarly, condition (iii) satisfies $|\pi(j)| \geq 1$ for $j \neq r$, and $\pi(r) = \emptyset$, while in (iv) we allow more roots to be present, as it was in (ii). This kind of topologies are not reconstructable with tree-specific algorithms, and thus only algorithms in [21, 20, 35, 34, 23] could be used for comparison.

The choice of these different type of topologies is not a mere technical exercise, but rather it is motivated, in our application of primary interest, by *heterogeneity of cancer cell types* and *possibility of multiple cells of origin*. In particular, type (ii) with respect to (i) and type (iv) with respect to (iii), are attempts at modeling independent progressions of a cancer via multiple roots. Clearly, these variations confound the inference problem further, since samples generated from such topologies will likely contain sets of mutations that are correlated but are pair-wise causally irrelevant – a well studied and widely discussed problem. Finally, note that, to generate synthetic data according to $(i - iv)$, the constraints on $\pi(\cdot)$ can be straightforwardly applied to the algorithm described below.

Generating synthetic data. Let n be the number of events we want to include in a DAG and let $p_{\min} = 0.05 = 1 - p_{\max}$, a *DAG without disconnected components* (i.e. an instance of type (iii) topology), maximum depth $\log n$ and where each node has at most w^* parents (i.e. $|\pi(j)| < w^*$, for $j \neq r$) is generated as follows:

1: pick an event $r \in G$ as the root of the DAG;
2: assign to each $j \neq r$ an integer in the interval $[2, \lceil \log n \rceil]$ representing its depth in the DAG (1 is reserved for r), ensure that each level has at least one event;
3: **for all** events $j \neq r$ **do**
4: let l be the level assigned to e;
5: pick $|\pi(j)|$ uniformly over $(0, w^*]$, and accordingly define $\pi(j)$ with events selected among those at which level $l - 1$ was assigned;
6: **end for**
7: assign $\alpha(r)$ a random value in the interval $[p_{\min}, p_{\max}]$;
8: **for all** events $j \neq r$ **do**

9: let y be a random value in the interval $[p_{\min}, p_{\max}]$, assign

$$\alpha(j) = y \prod_{x \in \pi(j)} \alpha(x)\,;$$

10: **end for**
11: **return** the generated DAG;

When an instance of type (iv) topology is to be generated, we repeat the above algorithm to create its constituent DAGs. In this case, if multiple DAGs are generated, each one with randomly sampled n_i events we require that $|G| = \sum n_i = n$. When instances of type (i) topology are required $w^* = 1$, and by iterating multiple independent sampling instances of type (ii) topology are generated. When required DAGs were sampled, these are used to generate an instance of the input matrix D for the reconstruction algorithms.

To account for noise in the data we introduce a parameter $\nu \in (0, 1)$ which represents the probability of each entry to be random in D, thus representing a *false positive* ϵ_+ and a *false negative* rate ϵ_-

$$\epsilon_+ = \epsilon_- = \frac{\nu}{2}\,.$$

Performance measures. We used synthetic data to evaluate the performance of CAPRI as a function of dataset size, ϵ_+ and ϵ_-.

In general, since our interest lies primarily in the causal structure underlying the progressive phenomenon of cancer evolution, we wish to measure the number of genuine claims inferred (*true positives*, TP), and the number of unidentified spurious causes (*false positives*, FP). Similarly, we will call *false negative* (FN) a genuine cause that we fail to recognize as causal and *true negative*, (TN) a cause correctly identified as spurious. With these measures we evaluated the rates of *precision* and *recall* as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{and} \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}\,.$$

The overall structural performance was measured in terms of the *Hamming Distance* (HD, [36]), the minimum-cost sequence of node edit operations (deletion and insertion) that transforms the reconstructed topology into the true ones (i.e., those generating data). This measure corresponds to just the sum of false positives and false negative and, for a set of n events, is bounded above by $n(n-1)$ when the reconstructed topology contains all the false negatives and positives.

Finally, to estimate reliable statistics, we use the following standard approach to assess the results. We generate, for each type of topology that we consider, 100 distinct progression models and, for each value of sample size and noise rate, we sample 10 datasets from each topology. Thus, every performance entry (Hamming, precision or recall) is the average of 1000 reconstruction results. This is the setting we use in most cases, unless differently specified.

5.1 Performance with different topologies and small datasets

Here we estimate the performance of CAPRI for datasets with sizes that are likely to be found in currently available cancer databases, such as The Cancer Genome Atlas, TCGA [37], i.e. $m \approx 250$ samples, and 15 events. The results are shown in Figure §7, for topologies (i) and (ii), and Figure §8, for topologies (iii) and (iv). There, we show all the results obtained by running the algorithm with bootstrap resampling, although results (data not shown) without this pre-processing leave the conclusions unchanged.

Results suggest a trend we may expect, namely that performance degrades as noise increases and sample size diminishes. However, it is particularly interesting to notice that, in various settings, CAPRI almost converges to a perfect score even with these small datasets. This happens for instance with type $(i - ii)$ topologies, where the Hamming distance almost drops to 0 for $m \geq 150$. In general, it is also clear that reconstructing forests is easier than trees, when the same number of events n is considered. This is a consequence of the fact that, once n is fixed, forests are likely to have less branches since every tree in the forest has less nodes. When reconstructing type $(iii - iv)$ topologies, instead, the convergence-speed of CAPRI to lower Hamming distance is slower, as one might reasonably expect. In fact, in those settings the distance never drops below 3, and more samples would be required to get a perfect score. We consider this to be a remarkable result, when compared to the worst-case Hamming distance value of $15 \cdot 14 = 210$. Panels of Figure §8 also suggest that disconnected DAGs are easier to reconstruct than connected ones, when a fixed number of events is considered. Similarly to the above, this could be credited to the fact that the size of the conjunctive claims is generally smaller, for fixed n. With respect to the precision and recall scores, one may note that CAPRI seems to be quite robust to noise, since the loss in the score-values appear nearly unaffected by any increase in the noise parameter.

5.2 Comparison with other reconstruction techniques

We compare now with state-of-the-art approaches introduced in §2, which, for the sake of clarity, are categorized as follows:

- ***Structural***: approaches include such algorithms as *Incremental Association Markov Blanket* (IAMB, [21]) and the *PC algorithm* [20], both subjected to log-likelihood maximization[11];
- ***Likelihood***: approaches encompass various maximum-likelihood approaches constrained by either the *Bayesian Dirichlet with likelihood equivalence* (BDE, [35]) or the *Bayesian Information Criterion* (BIC, [34]) scores;
- ***Hybrid***: approaches are mixed approaches as exemplified by *hidden Conjunctive Bayesian Networks* (CBN, [23]), and *Cancer Progression Inference with Single Edges* (see CAPRESE, [27]) which can be applied only to trees and forests.

For all the algorithms we used their standard R implementations: for IAMB, BDE and BIC we used package `bnlearn` [38], for the PC algorithm we used package `pcalg`, for CAPRESE we used `TRONCO` [39] (first release) and for CBN we used `h-cbn` [40].

Clearly, other algorithms exist in the literature, but we selected those which satisfied at least one of the following criteria: they seemed more effective in inferring causal claims (i.e., IAMB and PC), they regularize the Bayesian overfit (i.e., BDE and BIC), they assume a prior (i.e. BDE) or they were developed specifically for cancer progression inference (i.e., CBN and CAPRESE). Prominent among the ones, missing from this study, are the following: *Grow and Shrink* [41], which preliminary analysis have shown to be very similar to IAMB, and the *DiProg algorithm* [42], which unrealistically requires an input error rate to reconstruct a model; note that this kind of information is not generally available *a priori*.

[11] These are the classic versions of the algorithms discussed in §2, further subjected to log-likelihood optimization to assign a direction to all of the computed non-oriented edges (see the discussion on Markov equivalence classes in §2). This additional feature is necessary to permit a fair comparison against various structural approaches, which, otherwise, would be penalized with a worse Hamming distance, since these algorithms, in principle, can return non-oriented edges. Note that progression models, by their very nature, consist only of oriented structures.

Notice that we selected all the algorithms capable of inferring generic DAGs but CAPRESE [27], which can only be applied to infer trees or forests (i.e., type $(i - ii)$ topologies). In the literature there exist other approaches specifically tailored for such topologies, e.g., [28, 30], however since in [27] it is shown that CAPRESE is better than other approaches we can restrict our comparison. We place CAPRI in the *Hybrid* category though we clearly compare its performance with all the other approaches, with the aim of investigating which approach is more suitable to reconstruct the topologies we defined in the previous section.

The general trend is summarized in Figure §9, where we rank these algorithms according to the median performance they achieve, as a function of noise and sample size, and provide the parameters we used for comparison. In Figure §10 we compare CAPRI with the structural approaches (IAMB and PC). In Figure §11 we compare with the likelihood approaches (BIC and BDE) and, finally, in Figure §12 we compare with the hybrid ones. We remark that, because of the high computational cost of running CBNs (see the discussion in §2) the number of ensembles performed is 100 for CBNs, while it is 1000 for all other algorithms. Though this strategy provides less robust statistics for CBNs (i.e., less "smooth" performance surfaces), it is still sufficiently accurate to indicate the general comparative trends and relative performance efficiency.

5.3 Reconstruction without hypotheses: disjunctive causal claims

Recall that our algorithm expects as input all the hypothesized causal claims to infer more expressive logical formulas, i.e., claims with pure CNF formulas or even disjunctive claims over atomic events. Nonetheless, it is instructive to investigate its performance in two specific cases: namely, (i) without hypotheses ($\Phi = \emptyset$) and (ii) for datasets sampled from topologies with *disjunctive* causal claims.

To generate the input dataset we have to modify the generative procedure used for the other tests to reflect the switch from conjunctive to disjunctive causal claims. This task is actually rather simple, since we just change the labeling function α to account for the probability of picking any subset of the clauses in the disjunctive claim, and not picking the others. We use DAGs with 10 events and disjunctive causal claims with at most 3 atomic events involved, which is a reasonable size of a disjunctive claim, given the events considered. Clearly, this setting is generally harder than the one shown in Figures §10– §12, thus we expect performance to be somewhat inferior.

Here we compare CAPRI with all the algorithms used so far, and we show the result of this comparison in Figure §13, where $\Phi = \emptyset$, as noted earlier. The plot clearly confirms the trends suggested by previous analyses: namely, CAPRI infers the correct disjunctive claims more often than the others. Note also that the performance is measured on the reconstructed topology only, since, without input hypotheses, the algorithm evaluates only conjunctive claims, and does not allow different types of relations (e.g. disjunctions) to be inferred automatically. However, as anticipated, observed performance improvement is now much lower, and the Hamming distance fails to rise above 4. Furthermore, convergence to optimal performance was not observed for $m \leq 1000$, and it appears not to be reachable even for $m \gg 1000$ (at least so, when no hypotheses are used). It is also possible that, as n and the number of maximum disjunctive clauses increase, the result could be an even less satisfactory speed of convergence.

5.4 Reconstruction with hypotheses: synthetic lethality

We wondered whether CAPRI would be able to infer synthetic lethality relations, when these are directly hypothesized in the input set Φ. We started with a test of the simplest form: e.g.,

$$a \oplus b \rhd c,$$

for a set of events $G = \{a, b, c\}$ where we force progression from a to c to be preferential, i.e. it appears with 0.7 probability while b to c does so with only 0.3 probability. Despite this being the smallest possible causal claim, the goal was to estimate the *probability* of such a claim being robustly inferable, when $\Phi = \{a \oplus b \rhd c\}$, and its dependence on the sample size and noise. We measured the performance of all the algorithms, with an input lifted according to the claim so that all algorithms start with the same initial pieces of information. The performance metric estimates how likely an edge from $a \oplus b$ to c could be found in the reconstructed structures.

We show the results of this comparison in Figure §14. We note that CAPRI succeeds in inferring the synthetic lethality relation more than 93% of the times, irrespective of the noise and sample size used. More precisely, with $m \geq 60$ the algorithm infers the correct claim at any execution, thus suggesting that CAPRI, with the correct input hypotheses, is able to infer complicated claims, many of which could have high biological significance. Naturally, it would be reasonably expected that the performance of any of these algorithms would drop, were the target relations part of a bigger model.

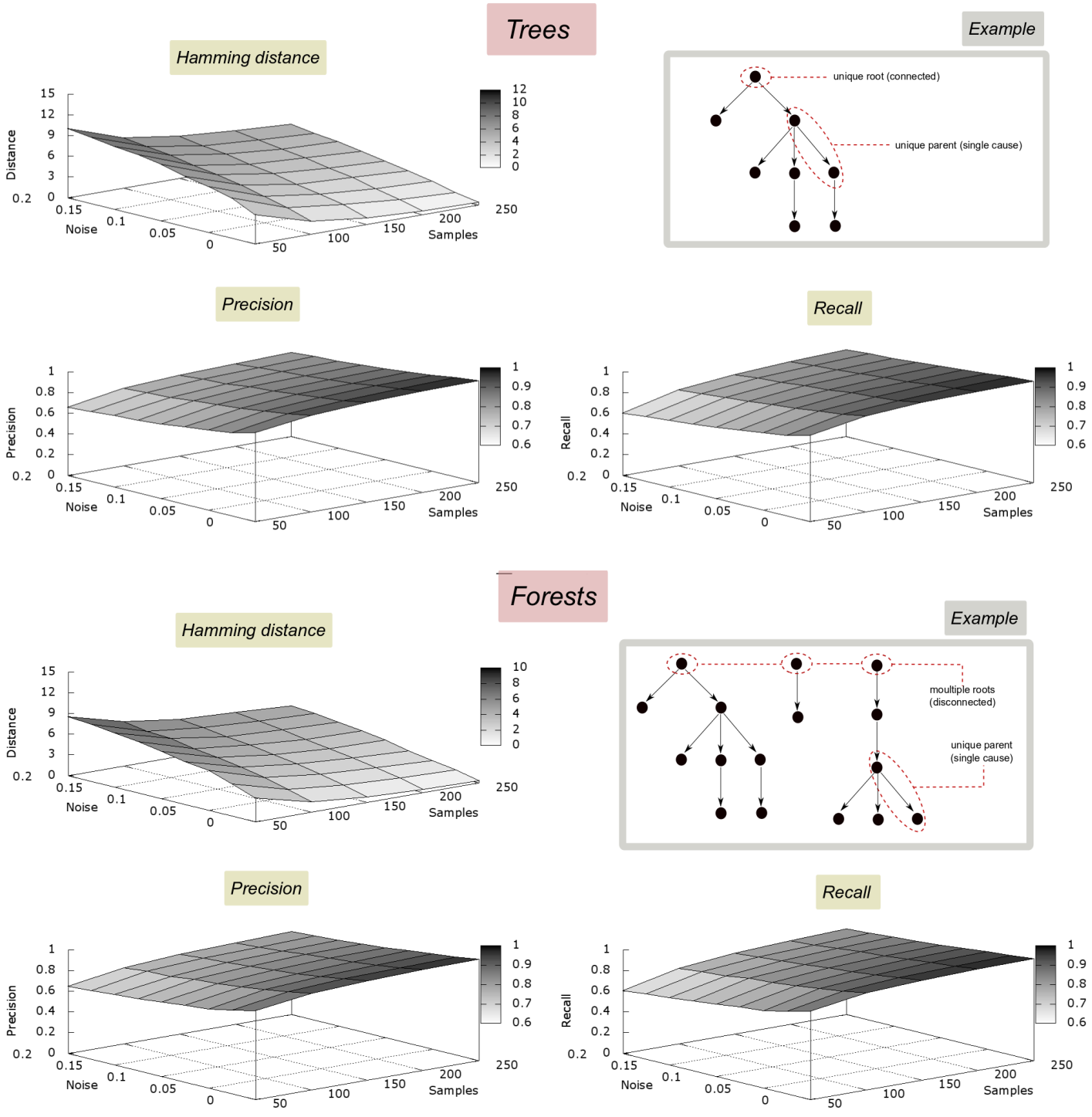


Figure 7: **Reconstruction of trees and forests with small datasets.** Hamming distance, precision and recall of CAPRI for synthetic data generated by trees (i.e., models with a single cause per event and a unique progression), in top panels, and by forests (i.e., models with a single cause per event but multiple independent progressions), in bottom panels. In both cases $n = 15$ events are considered, m ranges from 50 to 250 and the noise rate ranges from 0% to 20%. To have a reliable statistics we generate, for each type of topology, 100 distinct progression models and, for each value of sample size and noise rate, we sample 10 datasets from each topology. Thus, every performance entry is the average of 1000 reconstruction results. Notice that Hamming distance almost drops to 0 for $m \geq 150$ and that precision and recall decrease very little as noise increases.

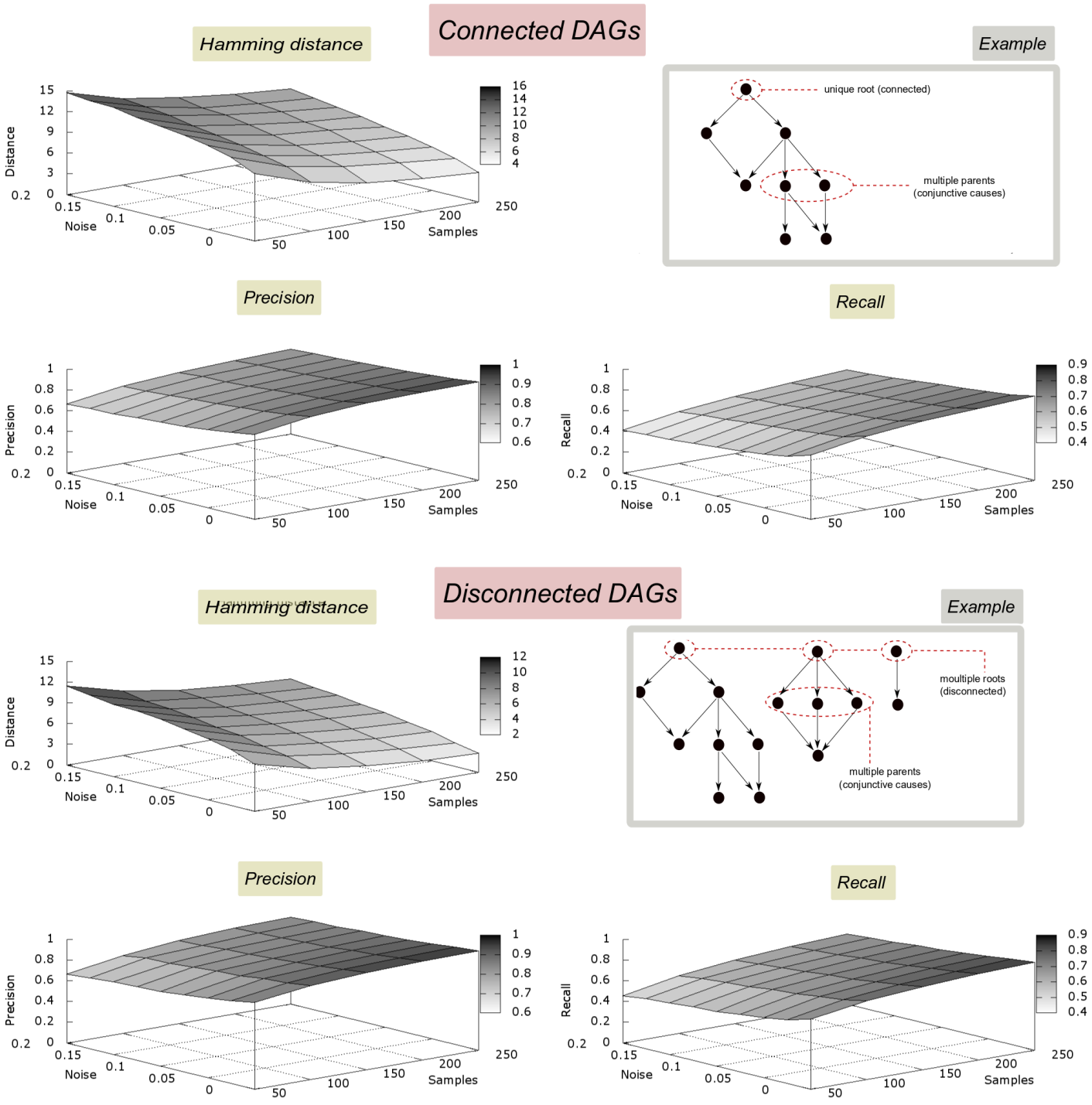
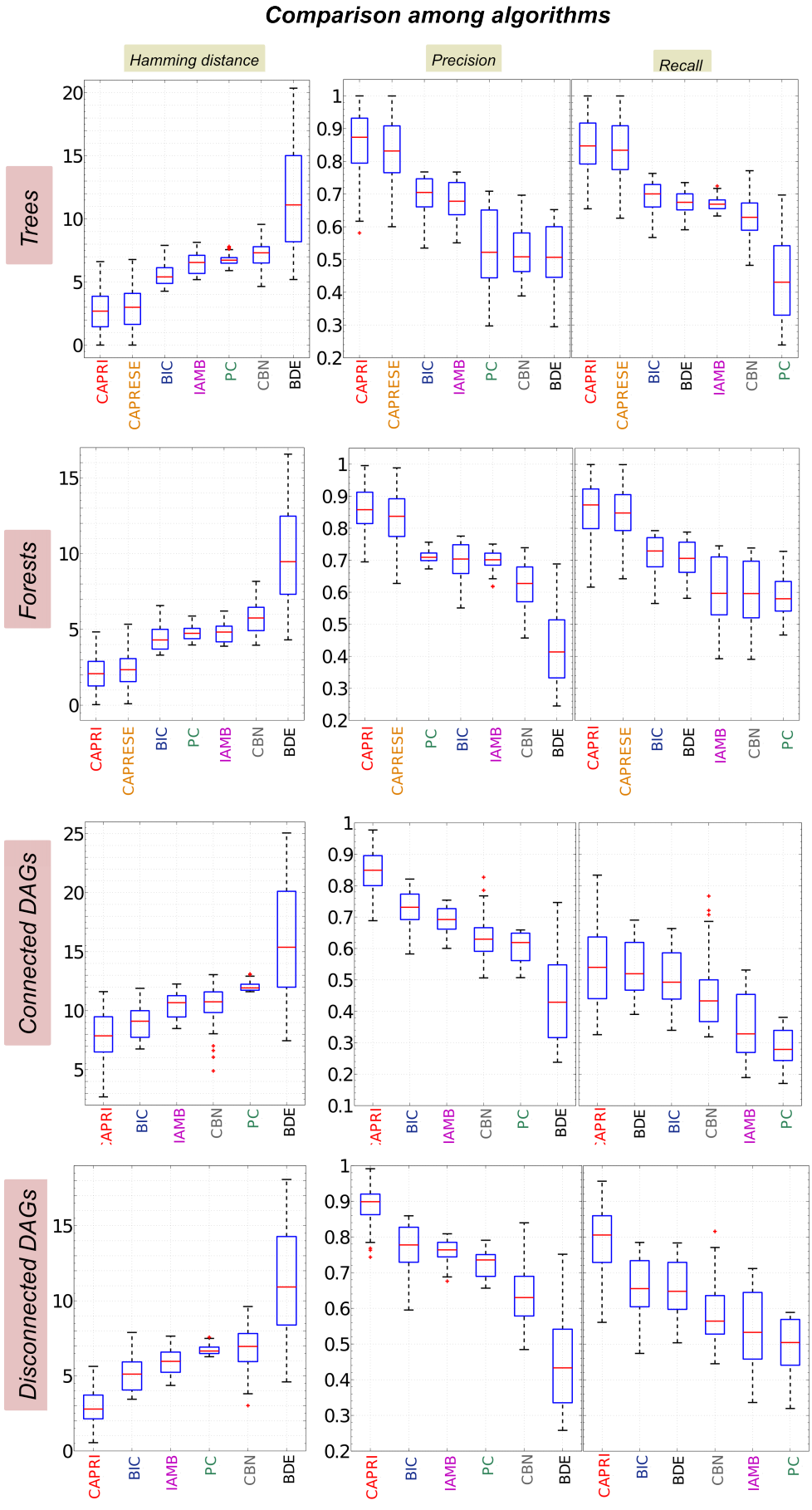


Figure 8: **Reconstruction of DAGs with small datasets.** Hamming distance, precision and recall of CAPRI for synthetic data generated by connected DAGs (i.e., models with either a single or conjunctive causes per event and a unique progression), in top panels, and by disconnected DAGs (i.e., models with either a single or conjunctive causes per event and multiple progressions), in bottom panels. In both cases the same parameters of Figure §7 are used ($n = 15$, $50 \leq m \leq 250$, $0\% \leq \nu \leq 20\%$ and every performance entry is the average of 1000 reconstructions). In this setting, which is harder than the one shown in Figure §7, Hamming distance does not reach values below 3 – a reasonably small number for our purposes – while precision and recall still decrease very little as noise increases.



Parameter values

n	*number of events*	10
m	*number of samples*	$[50, 1000]$
ν	*rate of false positives* ϵ_+ *and negatives* ϵ_-	$[0, 0.2]$ (0%-20% noise rate)
–	*ensemble size*	1000 (100 for CBN)

Figure 9: **Conjunctive causal claims: performance ranking.** We rank the algorithms we compared in Figure §10, §11 and §12 according to their performance for the parameters in the table. Rankings are divided according to the topology type and sorted according to the median performance.

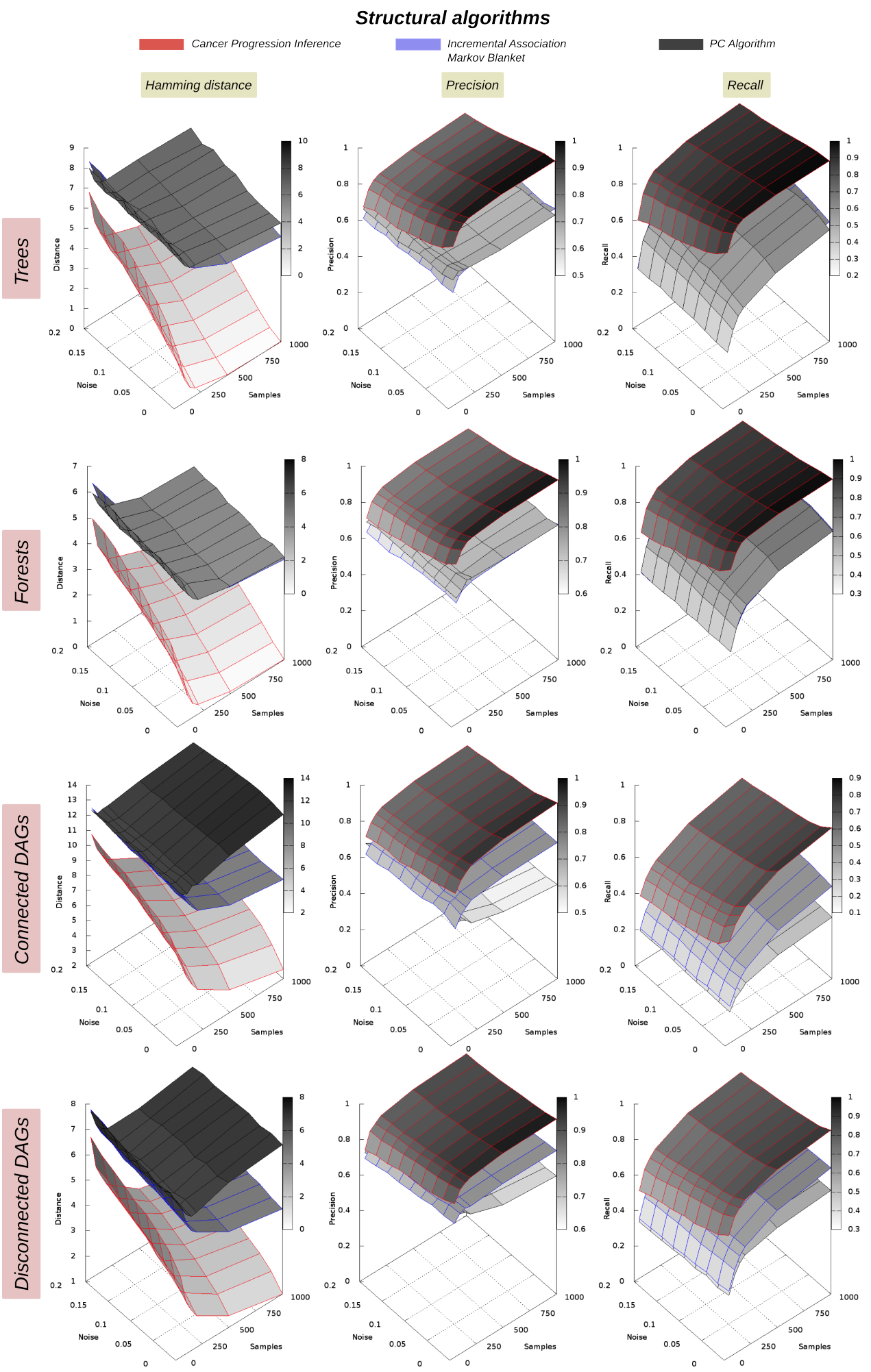


Figure 10: **Comparison with related works: structural algorithms.** We compare CAPRI, IAMB and the PC algorithm to infer *trees*, *forests*, *connected DAGs* and *disconnected DAGs* with the parameters described in Table §9. Average Hamming distance, precision and recall are shown.

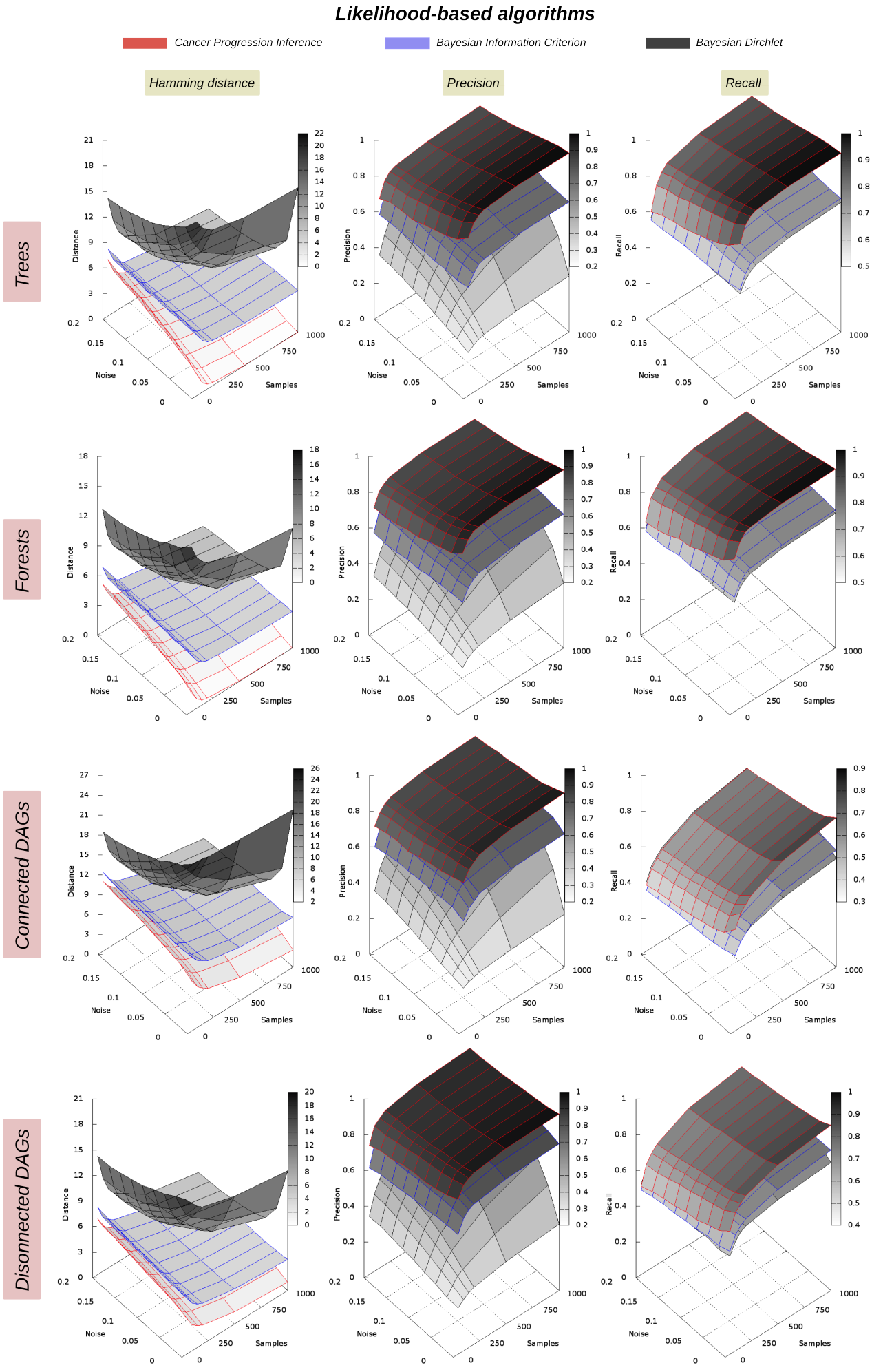


Figure 11: **Comparison with related works: likelihood-based algorithms.** We compare CAPRI, and that of likelihood-based methods based on BIC and BDE scores to infer *trees*, *forests*, *connected DAGs* and *disconnected DAGs* with the parameters described in Table §9. Average Hamming distance, precision and recall are shown.

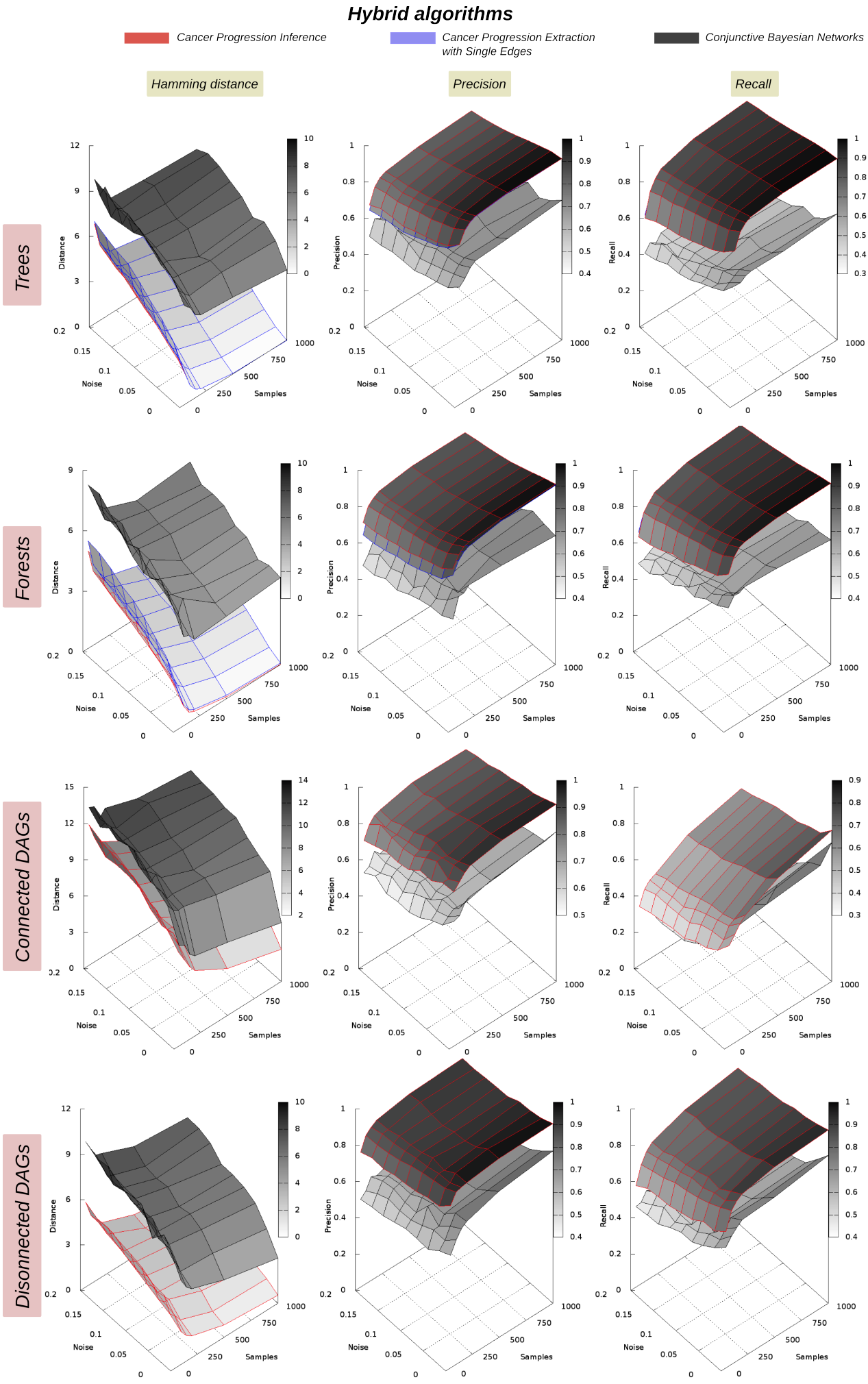


Figure 12: **Comparison with related works: hybrid algorithms.** We compare CAPRI, CBNs and CAPRESE to infer *trees*, *forests*, *connected* and *disconnected DAGs* with the parameters of Table §9 but, because of the computational cost of running CBNs with 100 annealing steps, we reduced the number of ensembles performed as: 100 for CBNs, 1000 for CAPRESE and, for CAPRI, 100 for DAGs and 1000 otherwise. Average Hamming distance, precision and recall are shown.

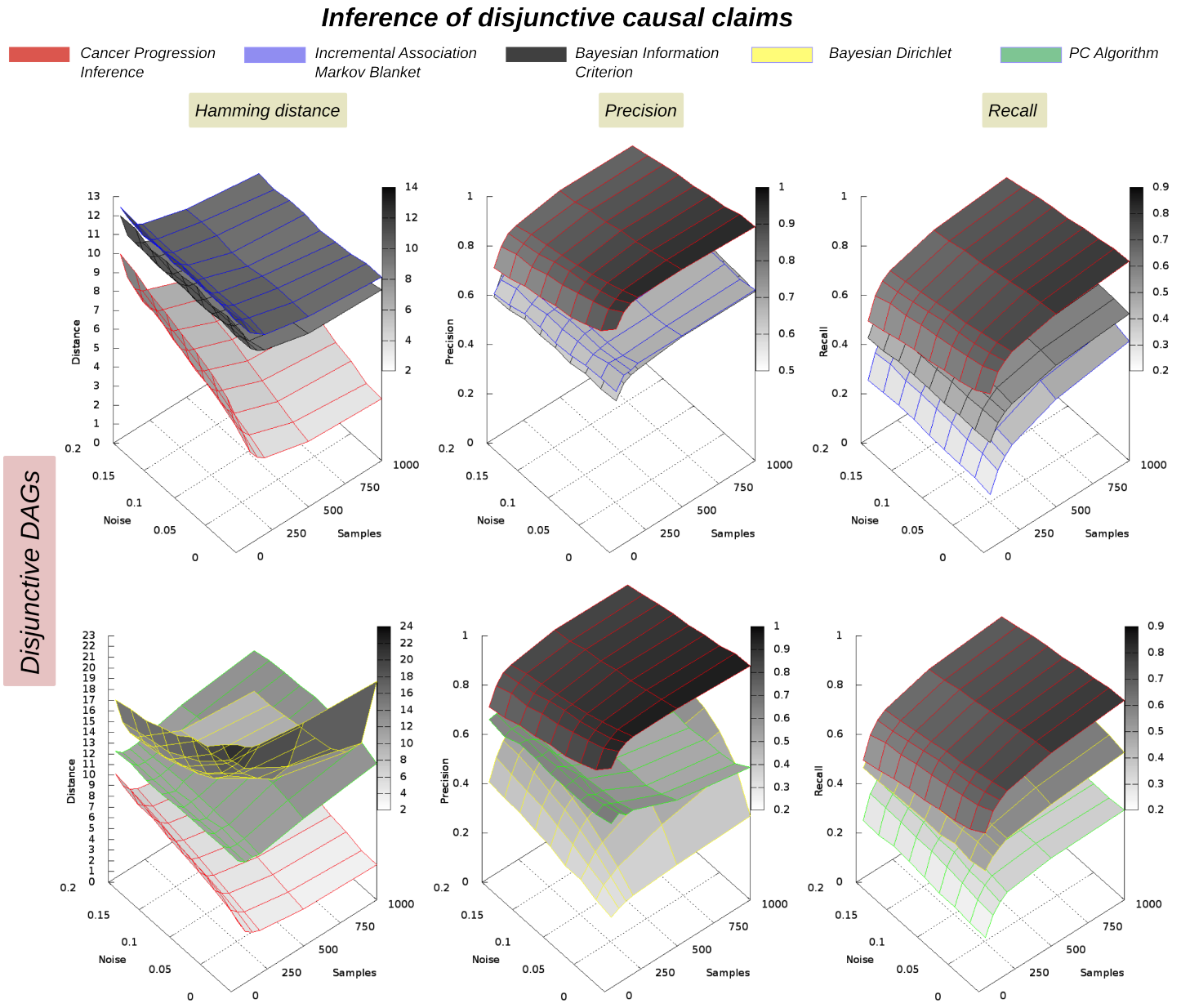


Figure 13: **Reconstruction of disjunctive causal claims with no hypotheses.** We compare CAPRI against all the algorithms to infer disjunctive causal claims. In top panel we show IAMB as the best structural algorithm, and the BIC score as the best among likelihood-based methods, according to Table §9. In bottom panel we compare the other algorithms. No hypotheses ($\Phi = \emptyset$) are given as input to CAPRI. Input data is generated by DAGs with 10 atomic events and disjunctive causal claims with at most 3 atomic events involved. Sample size ranges from 50 to 1000, noise rate from 0% to 20% and 1000 ensembles are generated for each configuration of noise and sample size. This setting is generally harder than the one shown in Figures §10– §12. Hamming distance, precision and recall are shown and confirm that disjunctions are harder than conjunctions to be inferred.

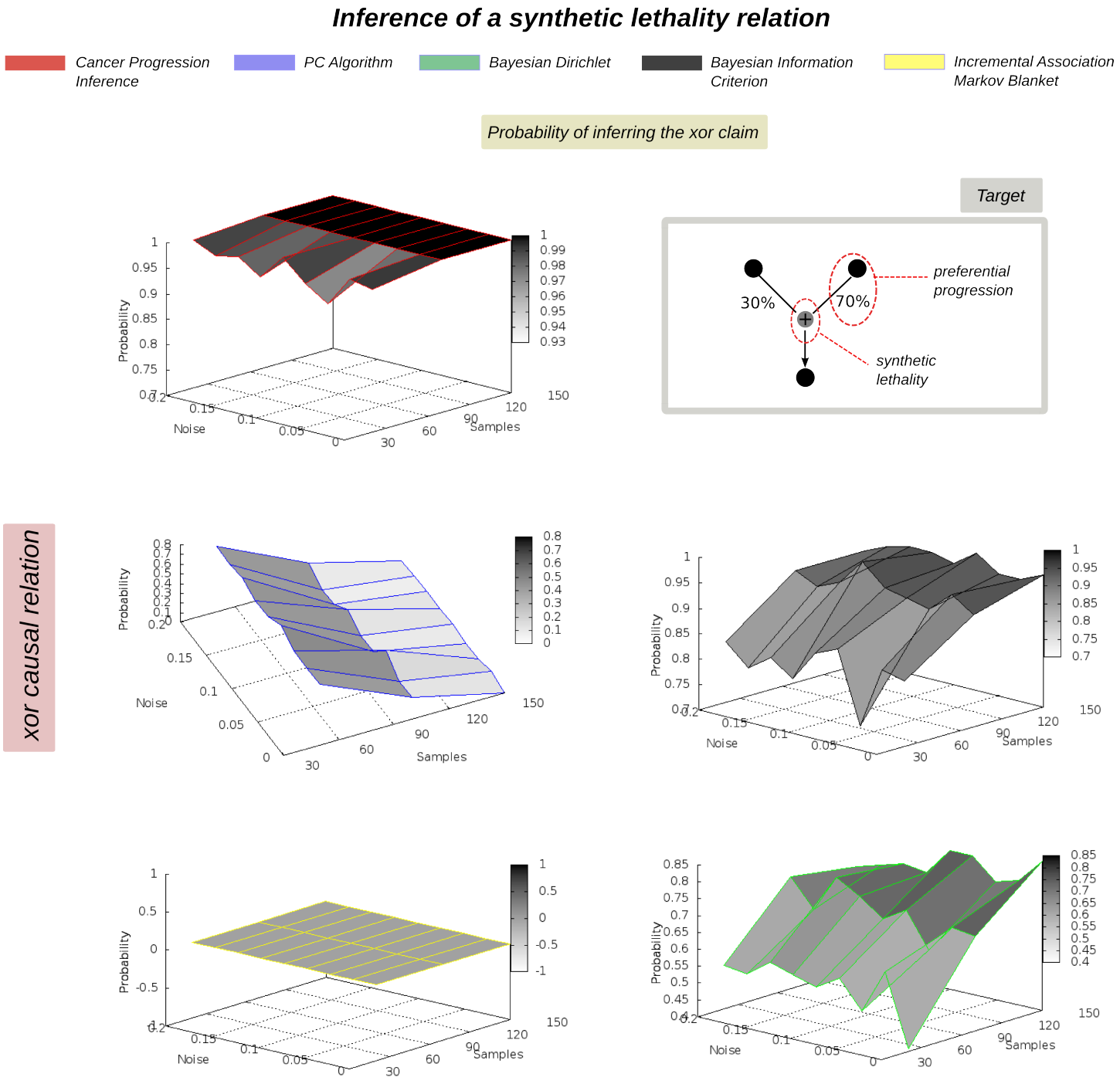


Figure 14: **Reconstruction with hypotheses: synthetic lethality.** We show the *average probability* of inferring a claim $a \oplus b \rhd c$ (*synthetic lethality*), when this is provided in the input set Φ. We show such a probability for CAPRI, the likelihood-based algorithms with BIC and BDE scores, and the structural IAMB and PC Algorithm. Data is generated from the model in the upper left panel (unbalanced "exclusive or" with a preferential progression), samples size ranges from 30 to 120, noise rate from 0% to 20% and 1000 ensembles are generated for each configuration of noise and sample size. Results suggest that a threshold level on the number of samples exists such that CAPRI infers the correct claim when $\Phi = \{a \oplus b \rhd c\}$. We executed all the algorithms with an input matrix lifted to contain the target claim.

6 Applications

Atypical Chronic Myeloid Leukemia. The p-values of both probability raising and temporal priority scores for the *Atypical Chronic Myeloid Leukemia* (aCML) dataset [43] are given in Supplementary File `pvalues-leukemia.xlsx`, whereas the result of the reconstruction with CAPRI is shown in the main text. Here, we show the results of reconstruction with other approaches, while delineating the differences in the structures reconstructed by CAPRI. We show in Figure §15 results of reconstruction with the structural algorithm *Incremental Association Markov Blanket with log-likelihood*, and the likelihood-based algorithm with *Bayesian Information Criterion* score. It is worth noting that only BIC infers the same relations on SETBP1 as those inferred by CAPRI. Somatic mutations considered here involve the following genes (see [43]): SETBP1, NRAS, KRAS, TET2, EZH2, CBL, ASXL1, IDH2, IDH1, WT1, SUZ, SF3B1, RUNX1, RBBP4, NPM1, JARID 2, JAK2, FLT3, EED, DNMT3A, EX23, CEBPA, EPHB3, ETNK1, GATA2, IRAK4, MTA2, CSF3R and KIT. In the plot we show only those events for which at least a causal claim was inferred.

Lung cancer. In Figure §16 we show a progression model of *Copy Number Variants* (CNVs) in lung cancer inferred with CAPRI from data published in [44]. The p-values of both probability raising and temporal priority scores are given in Supplementary File `pvalues-lung.xlsx`.

The dataset contains samples from 183 lung adenocarcinoma cases, and it was obtained by performing tumor/normal pairs with a combination of whole-exome sequencing or whole- genome sequencing, see [44] for a detailed commentary of data gathering. CNVs considered here involve the following genes (see Figure 2 in [44]): KRAS, EGFR, NKX2-1, MYC, MDM2, CCNE1, ERBB2, CCND1, TERT, CRKL, TP53 and CDKN2A. In the plot we show only those events for which at least a causal claim was inferred.

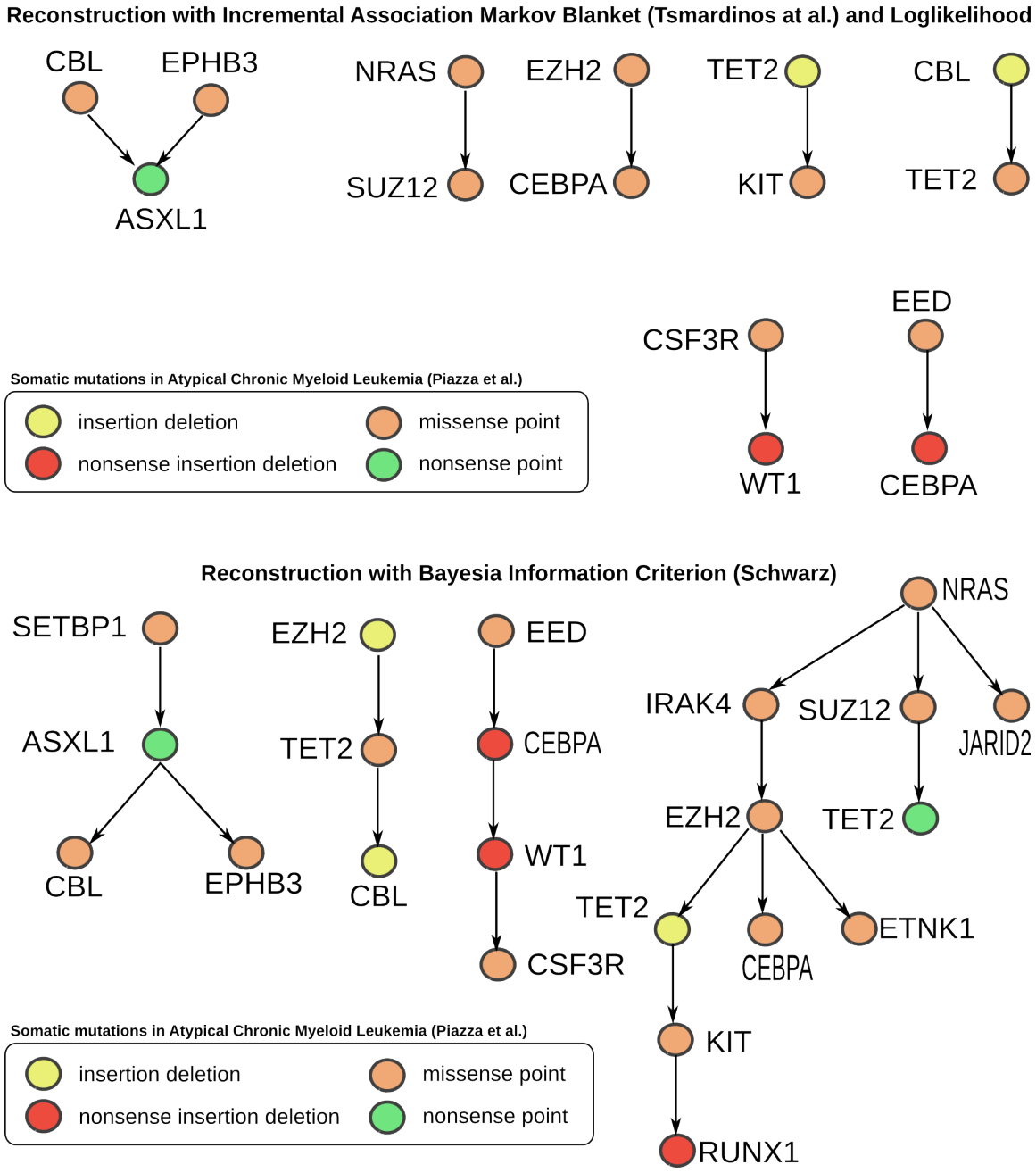


Figure 15: **Progression models of accumulating somatic mutations in aCML.** For the aCML dataset of [43] we show results of reconstruction with the structural algorithm *Incremental Association Markov Blanket with log-likelihood*, and the likelihood-based algorithm with *Bayesian Information Criterion* score.

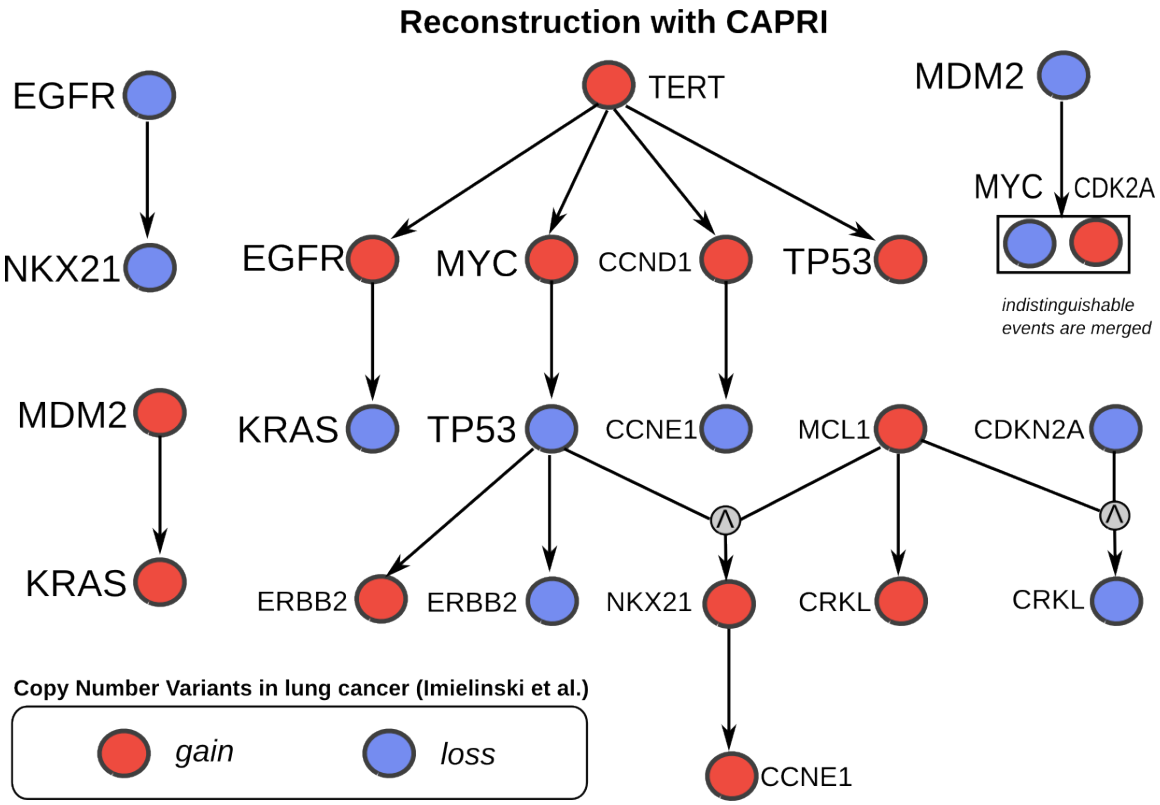


Figure 16: **Progression models of Copy Number Variants in lung cancer.** For the lung cancer dataset of [44] we show results of reconstruction with CAPRI.

References

[1] P. M. Illari, F. Russo, and J. Williamson, eds., *Causality in the Sciences*. Oxford University Press, 2011.

[2] C. Hitchcock, "Probabilistic causation," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), winter 2012 ed., 2012.

[3] J. B. Haldane, *The Causes of Evolution*. Princeton University Press, 1990.

[4] D. Hume, *An Enquiry Concerning Human Understanding*. 1748.

[5] H. Kyburg, "Discussion: Salmon's paper," *Philosophy of Science*, 1965.

[6] P. Suppes, *A Probabilistic Theory of Causality*. North-Holland Publishing Company, 1970.

[7] H. Reichenbach, *The Direction of Time*. University of California Press, 1956.

[8] N. Cartwright, *Causal Laws and Effective Strategies*. Noûs, 1979.

[9] B. Skyrms, *Causal Necessity*. Yale University Press, 1980.

[10] E. Eells, *Probabilistic Causality*. Cambridge University Press, 1991.

[11] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[12] P. Menzies, "Counterfactual theories of causation," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), spring 2014 ed., 2014.

[13] D. Lewis, "Causation," *Journal of Philosophy*, 1973.

[14] J. Woodward, "Causation and manipulability," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), winter 2013 ed., 2013.

[15] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[16] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[17] T. Verma and J. Pearl, "Equivalence and synthesis of causal models," in *Uncertainty in Artifical Intelligence Proceedings of the Sixth Conference* (M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, eds.), (San Francisco, CA, USA), pp. 220–227, Morgan Kaufmann, 1990.

[18] D. M. Chickering, "Learning bayesian networks is np-complete," in *Learning from data*, pp. 121–130, Springer, 1996.

[19] D. M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of bayesian networks is np-hard," *The Journal of Machine Learning Research*, vol. 5, pp. 1287–1330, 2004.

[20] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*, vol. 81. MIT press, 2000.

[21] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery.," in *FLAIRS Conference*, vol. 2003, pp. 376–381, 2003.

[22] A. M. Carvalho, "Scoring functions for learning bayesian networks," *Inesc-id Tec. Rep*, 2009.

[23] N. Beerenwinkel, N. Eriksson, and B. Sturmfels, "Conjunctive bayesian networks," *Bernoulli*, 2007.

[24] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel, "Quantifying cancer progression with conjunctive bayesian networks," *Bioinformatics*, vol. 25, no. 21, pp. 2809–2815, 2009.

[25] T. K. Moon, "The expectation-maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[26] S. Kirkpatrick, "Optimization by simulated annealing: Quantitative studies," *Journal of statistical physics*, vol. 34, no. 5-6, pp. 975–986, 1984.

[27] L. O. Loohuis, G. Caravagna, A. Graudenzi, D. Ramazzotti, G. Mauri, M. Antoniotti, and B. Mishra, "Inferring tree causal models of cancer progression with probability raising." Submitted for publication (available at arXiv.org)., 2013.

[28] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C. Papadimitriou, and A. Schäffer, "Inferring tree models for oncogenesis from comparative genome hybridization data," *Journal of Computational Biology*, 1999.

[29] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer, "Learning multiple evolutionary pathways from cross-sectional data," *Journal of Computational Biology*, 2005.

[30] A. Szabo and K. Boucher, "Estimating an oncogenetic tree when false negatives and positives are present," *Mathematical biosciences*, 2002.

[31] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM, 1982.

[32] B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2013.

[33] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[34] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, 1978.

[35] D. Heckerman, D. Geiger, and D. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, 1995.

[36] R. W. Hamming, "Error-detecting and error-correcting codes," *Bell System Technical Journal*, 1950.

[37] "The cancer genome atlas." `http://cancergenome.nih.gov/`, 2005.

[38] M. Scutari, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, 2010.

[39] "The `TRONCO` package for translational oncology." Available at standard R repositories.

[40] "Hidden conjunctive bayesian networks." `http://www.silva.bsse.ethz.ch/cbg/software/ct-cbn`.

[41] D. Margaritis, *Learning Bayesian Network Model Structure from Data*. PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA., 2003.

[42] H. S. Farahani and J. Lagergren, "Learning oncogenetic networks by reducing to mixed integer linear programming," *PLoS ONE*, 2013.

[43] R. Piazza, S. Valletta, N. Winkelmann, S. Redaelli, R. Spinelli, A. Pirola, L. Antolini, L. Mologni, C. Donadoni, E. Papaemmanuil, S. Schnittger, D.-W. Kim, J. Boultwood, F. Rossi, G. Gaipa, G. P. D. Martini, P. F. di Celle, H. G. Jang, V. Fantin, G. R. Bignell, V. Magistroni, T. Haferlach, E. M. Pogliani, P. J. Campbell, A. J. Chase, W. J. Tapper, N. C. P. Cross, and C. Gambacorti-Passerini, "Recurrent setbp1 mutations in atypical chronic myeloid leukemia," *Nature Genetics*, 2013.

[44] M. Imielinski *et al.*, "Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing," *Cell*, vol. 150, no. 6, pp. 1107–1120, 2012.

A Proofs

Here we collect all the proofs of the propositions and theorems stated in this document.

Proof of Propositions §1 and §2

Proof. These statements of Proposition §1 follow from the proof of Proposition §2, and we thus prove directly these statements.

- *(statistical dependence)* $\mathcal{P}(e \mid \varphi) > \mathcal{P}(e \mid \overline{\varphi}) \iff \mathcal{P}(e \wedge \varphi) > \mathcal{P}(e)\mathcal{P}(\varphi)$;
- *(monotonicity)* $\mathcal{P}(e \mid \varphi) > \mathcal{P}(e \mid \overline{\varphi}) \iff \mathcal{P}(\varphi \mid e) > \mathcal{P}(\varphi \mid \overline{e})$;
- *(ordering)* $\mathcal{P}(\varphi) > \mathcal{P}(e) \iff \dfrac{\mathcal{P}(e \mid \varphi)}{\mathcal{P}(e \mid \overline{\varphi})} > \dfrac{\mathcal{P}(\varphi \mid e)}{\mathcal{P}(\varphi \mid \overline{e})}$.

We prove the above statements starting from the same comments of [27] and by observing the algebraic subset relation subsumed by the hypothesis. □

Proof of Theorem §1

Proof. Recall that $k = |\Phi|$, $n = |G|$ and $D \in \{0,1\}^{m \times n}$, thus $D(\Phi)$ has $K = (n+k)m$ entries. We now analyze the complexity of CAPRI step-by-step.

- The cost of lifting depends on the input set Φ, if $\Phi = \emptyset$ it is $\mathcal{O}(1)$ both in time and space since $D(\emptyset) = D$.

  For non-empty sets, it requires evaluating $k \cdot m$ entries, after each claim $\varphi \rhd e$ is evaluated. Given that every φ has at worst n events included, its evaluation cost is at most $\mathcal{O}(n)$, even if lazy evaluation is performed. Thus, the cost of lifting is $\Theta(k \cdot m \cdot n)$, for a single bootstrap, which amplifies the bootstrap cost, as discussed in the previous section, in a multiplicative fashion. In terms of space, if $\Phi \neq \emptyset$ the overhead is $\Theta(K)$ if one copies D in $D(\Phi)$, $\Theta(km)$ otherwise.

- The cost of computing the parent function for the DAG requires a pair-wise calculation of the probabilistic scores, plus the cost of testing the $\sqsubseteq$ relation. Let $w = |N|$, where N is the set of nodes in the returned DAG. The score matrices for temporal priority and PR are $n \times w$, i.e. have columns for both atomic events and the disjunctive claims in the formulas of Φ, since we are disregarding causal claims of the form $\varphi_i \rhd \varphi_j$ and $a \rhd \varphi$ (differently, it would have been $w \times w$). Checking whether an atomic event is present in a disjunctive

claim is logarithmic in the size of the claim, if we lexicographically order its atomic events, thus bounded from above by $\log n$. Thus if we perform lazy evaluation for $\sqsubseteq$ the total number of comparison to select the parent function is at most

$$n[(n-1) + (w-n)\log n],$$

thus yielding a $\Theta(n^2)$ cost in time and space, if $w - n$ is small (it is 0 if $\Phi = \emptyset$), $\mathcal{O}(n(w - n)\log n)$ otherwise. In terms of space, the complexity is $\Theta(n[(n-1) + (w-n)])$, for a general Φ.

- As explained in CAPRI's definition, sometimes, albeit extremely rarely, a few extra operations might have to be performed when degenerate scores and loops are present. The procedure we suggested in CAPRI's definition requires sorting plus scan, thus its worst-case complexity is $\mathcal{O}(n \log n)$. Clearly, as this term is elided by the worst-case complexity of the steps discussed above, this unlikely scenario does not alter the complexity of the algorithm.
- Note that the cost of this analysis does not include the cost of BIC, as spelled out in the theorem statement.

The overall complexity follows, since:

- $\Phi = \emptyset$ then the major cost is that of evaluating $\mathcal{P}(\cdot)$ since usually $m \gg n$, thus $mn > n^2$. With regard to space, the only cost is that of book-keeping the scores.
- Let $m \gg n$ and $w - n > k$, in this case since $km \gg n$ and, under the mild assumption that $m > w$ and that k and $\log n$ are not relevant (in size) for m and w, then $km \gg (w-n)\log n$ which is the cost of lifting; thus is $\Theta(kmn)$ in time. Similarly, it follows that $mk \gg n[(n-1) + (w-n)]$.
- By computations similar to those carried out, it is indeed possible to see that $\mathcal{U}$, which is clearly finite since G is, grows *double-exponentially* in size with $|G|$ (i.e. the number of n-ary boolean functions, defined over the atomic events in any clause, possibly with negated literals), and thus the bound follows.

□

Proof of Theorem §2

Proof. We first prove the case with $\epsilon_+ = \epsilon_- = 0$, that is, the case where data have no noise. Some notations: (i) we denote with $\varphi \rhd e$ true claims (i.e. in $\mathcal{W}$), and (ii) with $\varphi^* \rhd e$ false ones. We divide the proof into several steps:

- First, we show that a prima facie DAG contains all the true causal claims, which is

$$\forall_{\varphi \rhd e \in \mathcal{W}}\ \pi(e) = \{\varphi\}\,.$$

By the event-persistence property usually assumed in cancer (fixating mutations are present in the progeny of a clone) the occurring times satisfy $t_\varphi < t_e$ which, in a frequentist sense, implies $\mathcal{P}(\varphi) > \mathcal{P}(e)$. In addiction, it holds by construction that $\mathcal{P}(\varphi \wedge e) = \mathcal{P}(e)$ when $\epsilon_+ = \epsilon_- = 0$, thus $\mathcal{P}(e \mid \varphi) = \mathcal{P}(e)/\mathcal{P}(\varphi)$, which is strictly positive since $\mathcal{P}(\varphi)$ and $\mathcal{P}(e)$ are, and that $\mathcal{P}(\overline{\varphi} \wedge e) = 0$, thus $\mathcal{P}(e \mid \overline{\varphi}) = 0$. Notice that $e \not\sqsubseteq \varphi$ by hypothesis.

- Now, we show that it might contain also spurious claims, which is

$$\exists_{\varphi^* \rhd e \notin \mathcal{W}}\ \pi(e) \subseteq \text{chunks}\,(\varphi^*) \cup \{\varphi^*\}\,.$$

These claims $\varphi^* \rhd e$ are of two types: sub-formulas spurious or topologically spurious (which include transitivities, as we may recall). For the former case note that

$$\forall_{\varphi \rhd e \in \mathcal{W}}\ \forall_{\hat{\varphi}^* \in \text{chunks}(\varphi)}\ \hat{\varphi}^* \rhd e \notin \mathcal{W},$$

but satisfies both temporal priority and probability raising. Also, consider any other $\hat{\varphi}^*_\star \sqsubseteq \hat{\varphi}^*$ and note that even this might satisfy both temporal priority and probability raising. For the latter case, it might be that there exists some other φ^* such that, it is positively statistically correlated to a real cause, and that might be prima facie as well.

Thus, for any $e \in G$ such that $\varphi \rhd e \in \mathcal{W}$

$$\pi(e) = \{\varphi\} \cup \mathcal{S},$$

where $\mathcal{S}$ is a set of spurious claims. We now examine the relation holding between the prima facie DAG and its modification performed via BIC. We denote these DAGs as $\mathcal{D}_{\text{pf}}$ and $\mathcal{D}_{\text{BIC}}$.

(i) First, we show that all true causal claims in $\mathcal{D}_{\text{pf}}$ are in $\mathcal{D}_{\text{BIC}}$, i.e.

$$\forall_{\varphi \rhd e \in \mathcal{W}}\ \pi_{\text{BIC}}(e) = \{\varphi\}\,.$$

Note that, although in general $\mathcal{P}(a \wedge b) \leq \min\{\mathcal{P}(a), \mathcal{P}(b)\}$, for the true claims following holds: $\mathcal{P}(\varphi \wedge e) = \mathcal{P}(e)$, when $\epsilon_+ = \epsilon_- = 0$; it is the maximum value for this joint probability, thus ensuring the maximum-likelihood fit. Thus the claim is maintained in $\mathcal{D}_{\text{BIC}}$.

(ii) Second, we need to show that if $\forall \varphi^* \rhd e \notin \mathcal{W}$ but present in $\mathcal{D}_{\text{pf}}$, there exists a claim $\varphi \rhd e \in \mathcal{W}$, which is present in $\mathcal{D}_{\text{pf}}$ and in $\mathcal{D}_{\text{BIC}}$ and any $\varphi^* \rhd e$ is not in $\mathcal{D}_{\text{BIC}}$.

Note that $\mathcal{P}(\varphi \wedge e) = \mathcal{P}(e)$, as above. Instead, $\mathcal{P}(\varphi^* \wedge e) < \mathcal{P}(e)$ since it is spurious, hence $\mathcal{P}(\varphi \wedge \varphi^* \wedge e) < \mathcal{P}(\varphi \wedge e)$, thus the likelihood fit of $\varphi \rhd e$ is maximal with respect to any of the claims $\varphi^* \rhd e$.

To extend the proof to $\epsilon_+ = \epsilon_- \in [0, 1)$ with uniform noise, it suffices to note that the marginal and joint probabilities change monotonically as a consequence of the assumption that the noise is uniform. Thus, all inequalities we used in the above proof still hold, which concludes the proof. □

Proof of Theorem §3

Proof. Consider the proof of the previous theorem. In this case, we are dealing with formulas such that $\text{chunks}\,(\varphi) \subseteq G$, i.e., formulas do not have any disjunctive component. All the derivations for Theorem §2 can be carried out in this context, notice that: formulas considered in step (i) of such a proof are those which are purely conjunctive and correctly inferred. Similarly, formulas in (ii) are those that screen off the false claims, but are incorrectly present in $\mathcal{D}_{\text{BIC}}$. □